

Detecting localized homogeneous anomalies over spatio-temporal data

Aditya Telang · P. Deepak · Salil Joshi ·
Prasad Deshpande · Ranjana Rajendran

Received: 17 January 2014 / Accepted: 18 June 2014 / Published online: 15 July 2014
© The Author(s) 2014

Abstract The last decade has witnessed an unprecedented growth in availability of data having spatio-temporal characteristics. Given the scale and richness of such data, finding spatio-temporal patterns that demonstrate significantly different behavior from their neighbors could be of interest for various application scenarios such as—weather modeling, analyzing spread of disease outbreaks, monitoring traffic congestions, and so on. In this paper, we propose an automated approach of exploring and discovering such anomalous patterns irrespective of the underlying domain from which the data is recovered. Our approach differs significantly from traditional methods of spatial outlier detection, and employs two phases—(i) discovering homogeneous regions, and (ii) evaluating these regions as anomalies based on their statistical difference from a generalized neighborhood. We evaluate the quality of our approach and distinguish it from existing techniques via an extensive experimental evaluation.

Responsible editor: Toon Calders, Floriana Esposito, Eyke Hüllermeier, Rosa Meo.

A. Telang (✉) · P. Deepak · S. Joshi · P. Deshpande
IBM Research, Bangalore, India
e-mail: aaditya.telang@in.ibm.com

P. Deepak
e-mail: deepak.s.p@in.ibm.com

S. Joshi
e-mail: saljoshi@in.ibm.com

P. Deshpande
e-mail: prasadesh@in.ibm.com

R. Rajendran
University of California, Santa Cruz, CA, USA
e-mail: ranjana.rajendran@gmail.com

1 Introduction

The growth in availability of geo-location sensing hardware and network connectivity has made it easier than ever to deploy sensors to monitor and aggregate information spanning large geographic regions over long periods of time. Interest in climate modeling and weather prediction has prompted deployment of hardware to sense temperature, pressure and humidity at very fine granularity. Urban planning and traffic management have sparked interest in monitoring flows in water supply and vehicular traffic to improve water management and schedule road works respectively. Regional disease incidence data may be analyzed across space and time to model and predict the spread of epidemics. In short, there has been tremendous growth in data having spatio-temporal characteristics. Given the scale and richness of such data, determining anomalous patterns is an interesting and important problem.

1.1 Motivating examples

We illustrate the notion of *anomalous patterns* with real-world examples.

Weather anomalies: Figure 1 represents a snapshot of the world map with temperature data at a specific time instance¹. The red and yellow regions represent the hot and cold extremes, whereas dark blue is used to color the oceans where no temperatures are recorded². One of the marked areas in the figure is the *Taklamakan desert*³ in the North Western China region. This region corresponds to a warm area of land encircled by mountains on three sides that are significantly colder at the time of the snapshot, and hence, is an obvious candidate for an *anomalous region*. Some other anomalies that are marked in the figure represent elevated cold regions in the Americas with warmer plains around them.

Twitter anomalies: Figure 2 plots the relative frequencies of the words *beer* and *church* in tweets⁴ originating from North America on July 4, 2012, the extremes represented by blue and red respectively. As may be expected, *church* peaks in the *Bible belt*⁵. However, some interesting anomalies can be observed wherein *church* tweets dominates a specific marked region in California, although the tweets in its neighbouring regions predominantly mention *beer*. Similar patterns can also be observed in the other marked regions of the mid-west.

The above examples illustrate that finding such regions that demonstrate significantly different behavior from its neighbors could be of interest for various application scenarios. In this paper, we propose an automated approach of discovering such anomalous regions irrespective of the underlying domain from which the data is recovered. Such anomalies may be verified or filtered using domain expertise later; for example,

¹ http://climate.geog.udel.edu/~climate/html_pages/download.html#ghcn_T_P2.

² In this paper, we extensively use color-based figures to illustrate the concepts of anomalies. Hence, we request the reader to refer to the electronic version or a colored printout of the paper for better readability.

³ http://en.wikipedia.org/wiki/Taklamakan_Desert.

⁴ <http://www.guardian.co.uk/news/datablog/2012/jul/04/us-fourth-july-twitter-beer-church>.

⁵ http://en.wikipedia.org/wiki/Bible_Belt.

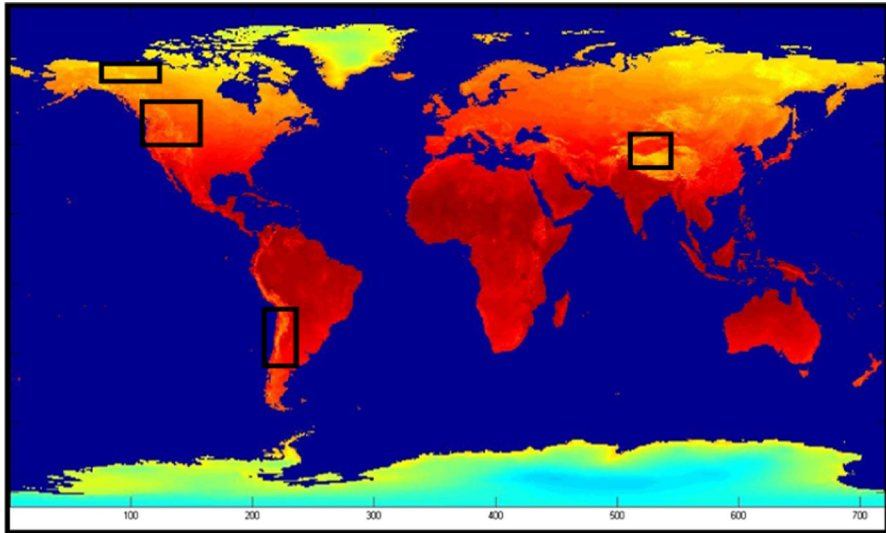


Fig. 1 Weather anomalies

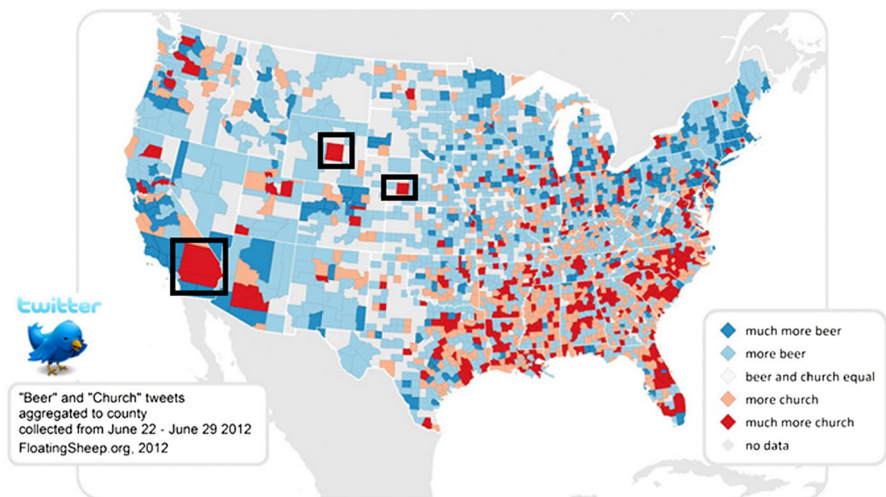


Fig. 2 Twitter anomalies

domain knowledge that *Taklamakan* is a desert helps in explaining the reason for this anomaly.

1.2 Characteristics of spatio-temporal anomalies

Given large-scale data with spatial, temporal as well as other parameters (*e.g.*, temperature, humidity, *etc.* associated with weather data), the goal of this work is to determine *spatio-temporal anomalous regions*. Formally, we define a spatio-temporal anomaly as

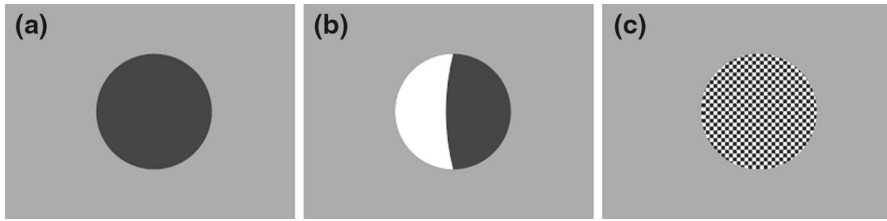


Fig. 3 Homogeneity example

a region which is – “homogeneous i.e., the values of data parameters being analyzed are consistent within the region, and statistically different from a local generalized neighborhood i.e., the data values within the region are significantly different from the ones in its neighborhood”. Let us analyze both aspects of this definition more clearly.

1.2.1 Homogeneity

Consider Fig. 3 that represents uni-variate spatial data (e.g., a single value such as temperature) that has been collected over points in a space, with the darkness of a point being directly proportional to the value of the reading. The black circular region in Fig. 3a is clearly anomalous due to having abnormally high readings compared to the surroundings. Our homogeneity criterion fails for the circular region in Fig. 3b due to the area having two (white and black) neighboring regions of contrasting readings. Intuitively, it may be argued that, analogous to the black region, the white region has significantly lower value as compared to its black and gray neighbors; hence, the two individual regions may be identified as separate homogeneous anomalies. In fact, some of the existing techniques for spatial outlier detection (Kou and tien Lu 2006) would identify both these regions as anomalous as they differ from their neighbors.

In contrast, Fig. 3c represents a case with a more or less uniform distribution of high and low values like the pattern in a checkerboard. Since the high and low values are mixed up, there are no sizable component homogeneous regions within the circular area. In contrast to existing techniques (Cheng and Li 2004; Shekhar et al. 2002) (which would classify each individual cell in the checkerboard as anomalous since it differs from its surroundings), we exclude regions such as those in Fig. 3c from being considered as candidates for anomalous regions due to the following reasons:

1. *Improbable occurrences:* We believe that uniformly scattered varying values such as the checkerboard pattern would occur in small regions. This is especially valid in the case of atmospheric data over regions in a geo-space and/or time. Furthermore, vast expanses of such regions are statistically improbable, and the ones observed following this pattern, would most likely be generated due to the possibility of noisy reading in difficult terrains. Such areas would be of little interest in the context of anomaly detection.
2. *Assumptions from statistical measures:* Statistical tools such as SaTScan (Kulldorff 1997) and its variants, use the assumption that data in the region in question is generated by a unimodal process (e.g., Poisson), and the notion of homogeneity

is consistent with this assumption. Additionally, under the spatial smoothness assumption, homogeneity may be used as a proxy for spatial coherence to limit the search space for anomalies.

3. *Preference to concise representations*: Anomalous homogeneous regions can be easily described using a concise description on the value-space. In contrast, non-uniform regions are difficult to describe intuitively. For example, Fig. 3c would be represented by a description such as $(temp > 0.8 \vee temp < 0.2)$, whereas a homogeneous region (Fig. 3a) would be easier to express using a single range (e.g., $(temp > 0.75)$).

1.2.2 Statistical difference from a local generalized neighborhood

An anomaly literally means something *out of the common*. We interpret this notion as being statistically different from a *local* and *generalized neighborhood*. Existing works on spatial anomaly detection (Kulldorff 1997; Patil and Taillie 2004; Neill and Moore 2004; Neill et al. 2005) typically classify a region as an anomaly if its analyzed parameters vary significantly from global parameters.

However, such an approach has significant drawbacks. For instance, a parameter like temperature is expected to increase gradually while moving inward from the periphery of a tropical desert, and high temperatures in the middle of the desert cannot be termed as anomalous despite being much higher than the average. On the other hand, high temperatures surrounded by a significantly colder regions is an interesting and uncommon occurrence, and would potentially need further inspection (e.g., hot springs embedded amongst regions of cold valleys). To the best of our knowledge, previous works have focused on *global divergences* and thus, are different from finding *locally* divergent anomalies.

Be that as it may, we do not classify every region which differs from its local neighborhood as anomalous. This is in stark contrast to some outlier-detection techniques (Kou and tien Lu 2006; Cheng and Li 2004; Shekhar et al. 2002) as well as image-segmentation techniques (Felzenszwalb and Huttenlocher 2004) that classify *every* homogeneous region as anomalous if it differs from its immediate neighbors. Instead we propose the notion of a *generalized neighborhood*. To illustrate this notion clearly, consider the circular region in Fig. 4a; its values are relatively higher com-

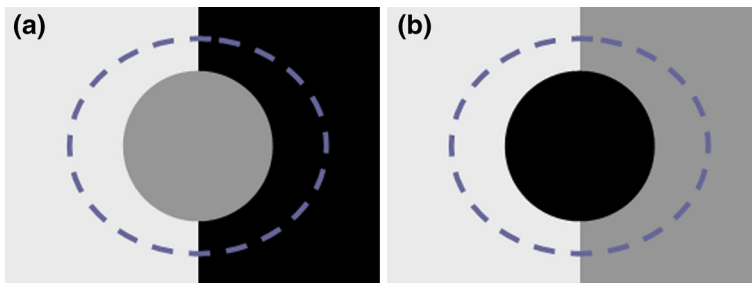


Fig. 4 Generalized neighborhood example: **a** transitional region, **b** anomaly

pared to the left side (and relatively lower than the right side). The ring around the circle bounded by the larger dotted circle forms its *neighborhood* region. Existing spatial-outlier detection as well as image segmentation techniques would consider the circular region as anomalous since it cannot expand to the left or to the right without violating the homogeneity condition. However, we propose that this region may not be considered anomalous since it simply represents an extended *transition* region of intermediate values between the low values from the left to the high values in the right.

On the other hand, Fig. 4b is clearly anomalous since its left and right neighborhoods contain low and intermediate values respectively, both contrasting well with the high values in the central circle. To re-emphasize, while the circles in Fig. 4a, b contrast well with their *local* neighborhood separately, the one in (a) does not contrast well with its *generalized* neighborhood since any measure of central tendency on the distribution of values in the neighborhood (comprising low and high values) would be quite close to the values within the circle.

1.3 Outline & contributions

The anomaly detection approach, proposed in this paper, comprises of two phases—(i) discover *homogeneous regions*, and (ii) evaluate such regions on their *statistical difference from the generalized neighborhood*. For phase one, we run a variant of agglomerative clustering (Lukasová 1979; El-Hamdouchi and Willett 1989) to generate homogeneous clusters (note that this approach can generate non-convex clusters as well). In the second phase, we filter out those clusters that are not sufficiently different from their generalized neighborhood using a statistical test, whereas those that survive are deemed to be *anomalous regions* (or anomalies).

The main *contributions* of this paper are:

- We introduce the novel problem of discovering *spatio-temporal anomalies* as *homogeneous* regions that are statistically different from their *local generalized neighborhood*. To the best of our knowledge, all previous works employ global statistics comparisons to ascertain anomalies.
- We present a two-phase approach for discovering anomalies and establish through a user study that our technique outperforms the previous methods in identifying intuitive anomalies more accurately.

The rest of the paper is organized as follows: In Sect. 2, we survey the related work. Sect. 3 formally defines the problem, Sect. 4 explains our approach for the same, and Sect. 5 details the results of our experimental evaluation along with a brief analysis of these results. Finally, Sect. 6 concludes the paper with directions for future work.

2 Related work

2.1 Overview of related work

Our problem of anomaly detection can be seen as a specialization of the general problem of *identifying data with divergent behavior*. The high-level goal of characterizing

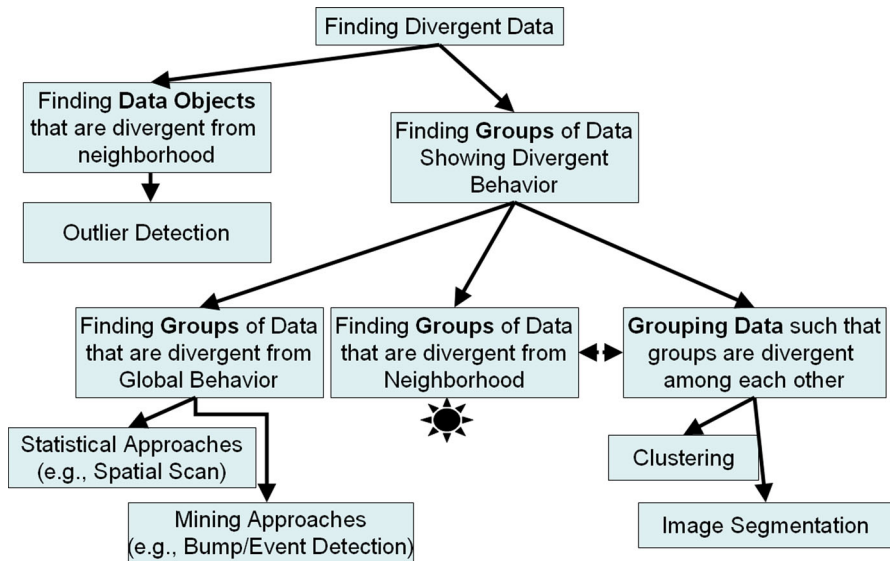


Fig. 5 Taxonomy of approaches for finding divergent data

data with respect to behavioral differences has been addressed in several different tasks ranging from outlier detection to clustering. An overview of techniques that can be used to find data with divergent behavior appears in Fig. 5. Techniques can be broadly classified as to whether they seek to estimate divergent behavior at the individual data object level or at the level of groups of data. Outlier detection techniques operate at the object level, wherein they quantify each data object w.r.t their difference from (most usually) the local neighborhood. Among the techniques that seek to identify groups of data objects, there are two types of approaches; (1) finding data groups that are divergent from global behavior (i.e., behavior estimated at the level of the whole dataset), and (2) partitioning the whole dataset into groups such that the groups are divergent from each other. Statistical approaches such as scan statistics and mining approaches such as spatial event detection are of the first kind, whereas clustering and image segmentation approaches fall into the second category. We address the highlighted problem of finding groups of data that are divergent from the neighborhood (the generalized local neighborhood, in particular). As indicated by the double-edged arrow in the figure, *grouping* techniques such as clustering and image segmentation are related to our problem since ensuring divergence between groups automatically ensures some divergence from the neighborhood due to the neighborhood itself being part of another group(s). We now detail the differences of our work from the various groups of techniques in separate sub-sections herein.

2.2 Outlier detection approaches

Outlier detection is the problem of quantifying, for any object/observation, the inconsistency (i.e., outlierness) between itself and the remainder of the data; the objects

with the highest outlierness are then deemed to be outliers. A recent work (Schubert et al. 2014) surveys the different outlier detection techniques along with those that treat spatial attributes specially. Spatial outlier detection has been extensively studied in the geospatial and geosciences community. However, as pointed out in Sect. 1, outlier detection techniques (Kou and tien Lu 2006; Cheng and Li 2004; Shekhar et al. 2002) perform a single point/observation level estimation, and hence, differ from the basic definition of an anomalous region as proposed in this paper. For instance, given a spatial grid, these techniques will consider a grid cell as an outlier if its values are divergent from its immediate neighbors. Hence, these methods will end up typically classifying all the individual cells in the checkerboard pattern (in Fig. 3c) as outliers; which, as argued in Sect. 1 differs from our definition of an anomalous region. Furthermore, none of these works adopt the notion of a generalized neighborhood, which is a major distinguishing component of our work. In addition, since these techniques only consider *individual* grid cells (Kou and tien Lu 2006; Cheng and Li 2004) or individual graph edges (Shekhar et al. 2002) as candidates for outliers, the final set of outliers detected are neither homogeneous nor arbitrary-shaped.

2.3 Statistical approaches for identifying globally divergent groups of data

The problem of finding globally divergent regions has been extensively studied in the statistics community, where sampling regular regions such as circles followed by a likelihood ratio test to assess divergence (Kulldorff 1997) has been a popular approach. Spatial scan statistics have been refined to identify arbitrary-shaped regions in methods such as ULS Scan (Patil and Taillie 2004) whereas index-based (Neill and Moore 2004) and simulated annealing based region growing approaches (Duczmal 2004) have also been proposed towards the same problem. In Tango and Takahashi (2005), authors argue that allowing for unconstrained arbitrary regions can sometimes be bad, and provide a method to restrict the shape to avoid *peculiar* regions where faraway spaces are brought together into the same region. Additionally, new types of statistical tests such as the bayesian spatial scan statistic (Neill et al. 2005) have been proposed and shown to help find globally divergent regions faster. However, as pointed out in Sect. 1 and to the best of our knowledge, previous work on finding spatial events has only focused on *global divergences* and are thus different from our problem of finding locally divergent anomalies.

2.4 Mining approaches for identifying globally divergent groups of data

The mining community has also addressed the problem of identifying globally divergent behavior in the context of detecting spatial events *i.e.*, those areas that differ from average behavior of the entire space under consideration. With the average behavior learnt from across the dataset (*i.e.*, *global behavior*) in a pre-processing phase, the spatial event detection problem could be seen as searching for those areas where the local behavior is divergent from the global (*i.e.* *globally divergent* regions). This could be done by hierarchically drilling down towards globally divergent areas in a top-down fashion (Friedman and Fisher 1999), or by means of a bottom-up approach

where seed objects whose neighborhoods display divergent behavior are aggregated to form globally divergent areas (Ester et al. 1996). Once the global behavior is learnt, a candidate region may be scored by assessing its behavior, and comparing against the learnt global behavior.

2.5 Clustering

Among the most popular techniques to group data into homogeneous clusters (that are mutually divergent) are *clustering* techniques (Jain et al. 1999) that seek to minimize the intra-cluster distance. However, general clustering techniques usually do not differentiate between spatial and non-spatial (e.g., temperature) attributes; thus, application of clustering to a dataset of sensors could group sensors with very divergent readings together if they are very close in space. In particular, the uniform treatment makes it impossible to identify clusters that are spatially connected while being homogeneous on the non-spatial attributes (since the difference in criteria entails a requirement of differential treatment). Though techniques such as ST-DBSCAN (Birant and Kut 2007) propose to treat spatial and non-spatial attributes differentially, the clusters in the output are not necessarily contiguous in space since spatial proximity can still offset for non-spatial homogeneity. An adaptation of hierarchical agglomerative clustering (we call it *HAC-A*) would merge the pair of *adjacent* clusters that are closest on the readings attribute and discover homogeneous and arbitrary shaped regions. However, they may not necessarily differ from the *generalized local neighborhood* since clusters consider only homogeneity and are oblivious to the contrast with the local neighborhood. Spatio-temporal clustering (Kisilevich et al. 2010), the field relating to clustering as applied to observations that have spatial and temporal attributes, have mostly focused on moving object data such as trajectories where sequences of spatio-temporal points are considered as single objects to be clustered. Additionally, there have been many special-purpose algorithms that seek to identify specific patterns; for example, cyclone trajectories could be detected (Stolorz et al. 1995) as sequences of low-pressure spatio-temporal points that are in temporal sequence and coherent with extrinsic data such as windspeed. Another work deals with clustering cellular towers (Reades et al. 2007) using just the load information (i.e., number of calls passing through it), where each cellular tower has a set of features, each indicating the load factor during a specific time window.

2.6 Image segmentation

Image segmentation techniques (e.g., blob detection), widely studied in the computer vision community, employ histograms (Ohlander et al. 1978; Bonnet et al. 2002), graph partitioning (Shi and Malik 2000; Grady and Schwartz 2006) and region growing (Revol and Jourlin 1997; Fan et al. 2001) to identify regions with homogeneous coloring. These methods are more relevant to our problem than clustering since the color attribute could be conveniently replaced by other parameters (like temperature, tweets, etc.). However, like clustering, they too do not use any generalized neighborhood comparison in prioritizing regions.

Table 1 Related work summary

| Technique | Homogeneity | Generalized local neighborhood | Arbitrary-shaped |
|---|-------------|--------------------------------|------------------|
| Spatial outlier detection (e.g., Shekhar et al. 2002; Kou and tien Lu 2006) | × | × | × |
| Spatial scan statistic Kulldorff (1997) | × | × | × |
| ULS scan Patil and Taillie (2004), FlexiScan Tango and Takahashi (2005) | × | × | ✓ |
| Spatial event detection (e.g., Friedman and Fisher 1999; Ester et al. 1996) | × | × | ✓ |
| HAC-A (Ref. Sect. 2) | ✓ | × | ✓ |
| Image segmentation (e.g., Arbelaez et al. 2011) | ✓ | × | ✓ |

2.7 Summary of related work

To summarize the discussion, we present a comparison of some techniques techniques in literature with respect to our three criteria for anomaly detection in Table 1. While HAC-A and image segmentation techniques make use of two of our three criteria, they do not exploit the generalized local neighborhood. Thus, these could be potential replacements to the first-phase of discovering homogeneous regions, in our approach, as outlined in Sect. 1.3. Nevertheless, we will compare our technique against several of these approaches in the experimental analysis.

3 Problem formulation

Let $\mathbf{S} = \{C_1, C_2, \dots, C_{x \times y \times t}\}$ be a spatio-temporal gridded cube over the spatial (x,y) and temporal (t) dimensions. We use the single suffix notation (e.g., C_i) instead of representing the cubes as-is (e.g., $C_{(i,j,k)}$) for simplicity. Let $\mathbf{A} = \{A_1, A_2, \dots, A_q\}$ be a set of attributes over which spatio-temporal anomalies will be defined. In the context of weather data, these attributes could be temperature, air-pressure, humidity, air-density and so on. Every $C_i \in \mathbf{S}$ then represents a vector of the form— $\{v_i^1, v_i^2, \dots, v_i^q\}$, wherein v_i^k represents a value in the domain of attribute $A_k \in \mathbf{A}$. For instance, consider a sample 8×8 grid⁶ shown in Fig. 6 defined over a single attribute of temperature such that each cell represents a specific attribute value associated with the cell region. In the figure, white cells represent temperatures below 10° whereas the gray ones are above 25° .

Now, consider a set $S_A : \{C_1, C_2, C_3, \dots, C_s\}$, such that $S_A \subset \mathbf{S}$. We classify S_A as a *spatio-temporal anomaly* if it satisfies the:

⁶ For sake of clarity, we illustrate a spatial grid; however, the formulation is extendible to the temporal dimension.

Fig. 6 Example: problem definition

| C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 |
|-----|-----|-----|-----|-----|-----|-----|-----|
| C9 | C10 | C11 | C12 | C13 | C14 | C15 | C16 |
| C17 | C18 | C19 | C20 | C21 | C22 | C23 | C24 |
| C25 | C26 | C27 | C28 | C29 | C30 | C31 | C32 |
| C33 | C34 | C35 | C36 | C37 | C38 | C39 | C40 |
| C41 | C42 | C43 | C44 | C45 | C46 | C47 | C48 |
| C49 | C50 | C51 | C52 | C53 | C54 | C55 | C56 |
| C57 | C58 | C59 | C60 | C61 | C62 | C63 | C64 |

Temp < 10
 Temp > 25

1. *Spatio-temporal connectedness condition*: S_A is a spatio-temporally connected region; i.e., when a graph is constructed from S_A where cells are nodes, and edges are induced between all pairs of neighboring cells in S_A , we require that any pair of nodes $\{C_i, C_j\} \in S_A$ should be reachable through a sequence of edges.
2. *Homogeneity condition*: For the individual distribution of values for each distinct attribute across all elements in S_A , i.e., $\{v_1, v_2, \dots, v_s\}$, we require that a dispersion measure $Dispersion(\{v_1, \dots, v_s\})$ evaluate to not more than a threshold τ . Among various options for quantifying dispersion (e.g., Gini co-efficient (Ceriani and Verme 2012), Quartile co-efficient (Bonett 2006) and reciprocal of entropy), we choose the Gini co-efficient in our method.
3. *Neighborhood heterogeneity condition*: We define a generalized neighborhood region for S_A as comprising of all cubes from the space S that have at least one cube from S_A at a spatio-temporal distance not more than ρ :

$$\mathcal{N}_{S_A} = \{s | s \in S : \exists s' \in S_A, dist(s, s') \leq \rho\}$$

where $dist(., .)$ is measured by a popular distance metric such as Chebyshev distance⁷. Informally, \mathcal{N}_{S_A} defines a region of width ρ enveloping the region defined by S_A . Our neighborhood heterogeneity condition requires that the values in the cubes within S_A be sufficiently different from those in \mathcal{N}_{S_A} , denoted as $\{v'_1, \dots, v'_{|\mathcal{N}_{S_A}|}\}$. Specifically, we prefer that the value of $Stat(\{v_1, \dots, v_s\}, \{v'_1, \dots, v'_{|\mathcal{N}_{S_A}|}\})$ be maximized where $Stat(., .)$ is any measure (such as likelihood ratio test (LRT) (Mood et al. 1963), Chi-squared test and paired t test (Mankiewicz 2000)) for estimating statistical divergence between distributions. In this paper, we choose to use the LRT test.

For the spatial grid in Fig. 6, the set $\{C_{13}, C_{19}, C_{20}, C_{21}, C_{28}, C_{29}, C_{37}, C_{38}\}$ represents an anomaly (S_A). The set of cells are connected as may be seen from the figure, thus satisfying the connectedness condition. These cells all form high-temperature cells (gray color), and are hence homogeneous too. The gray region is surrounded by white cells of low temperature, that form \mathcal{N}_{S_A} ; the neighborhood heterogeneity condi-

⁷ http://en.wikipedia.org/wiki/Chebyshev_distance.

tion would also be met for this region since all the cells within it are high temperature cells, and those outside are all low-temperature, ensuring high statistical divergence.

It must be noted that so far we have defined the notion of a spatio-temporal anomaly only in the context of a gridded cube. However, this formulation can be intuitively extended to any dataset where the neighborhood relation is well-defined. For example, road networks can be modeled by considering roads incident on the same intersection, as being neighbors.

4 The anomaly detection approach

Given a spatio-temporal region with well-defined neighbors and well-defined parameter values within each region, our algorithm (Algorithm 1) detects anomalies using a two-step process.

4.1 Cluster formation

We start by marking all cells in the grid as *unclustered* (Line 2 in Algorithm 1). From these unclustered cells, we pick an arbitrary cell (Line 4) and try to grow it to form a homogeneous cluster. Towards this, at any step of the merging process, the cluster is compared to each neighboring cell (that is adjacent to at least one cell in the cluster),

Alg. 1 *Anomaly detection*

Input. Grid G with input values

Input. gini indexing threshold τ

Input. LRT statistic threshold γ

Output. Set of anomalies A

/ Cluster Formation Phase */*

```

1.  $Clusters \leftarrow \{\}$ 
2.  $Unclustered \leftarrow \{c | c \in G\}$ 
3. while  $|Unclustered| > 0$  do
4.    $c = \text{next cell from } Unclustered$ 
     acc to chosen ordering
5.    $C \leftarrow \{c\}$  // cluster initialization
6.   while true do
7.      $c' \leftarrow \arg \min_{c \in neighbor(C)} (gini(C \cup \{c\}))$ 
8.     if  $(gini(C \cup \{c'\}) \leq \tau)$ 
9.        $C \leftarrow C \cup \{c'\}$ 
10.    else
11.       $Clusters = Clusters \cup C$ 
12.       $Unclustered = Unclustered - C$ 
13.      break
14.    end if
15.  end while
16. end while
```

/ Anomaly Detection Phase */*

```

17.  $A = \{C | C \in Clusters \wedge LRT(C) > \gamma\}$ 
18. return  $A$ 
```

and the one whose merger would result in the least dispersion value for the cluster is chosen and added to the current cluster (Line 7). However, the merger is affected iff the merged cluster has a dispersion value within τ (Line 9). When no more mergers can be performed to grow the cluster, we include it in the list of candidate clusters (Line 11), and mark the component cells as *clustered* (Line 12). Another unclustered cell is then chosen as a seed, and this process is repeated until all cells are clustered. The seeds may be chosen according to some pre-determined ordering of cells (*e.g.*, Z-order, or row-major order).

The dispersion of a set of cells is computed as the dispersion in the distribution of the readings (*e.g.*, temperature, pressure, or any sensor reading) within those cells; in particular, the spatial or temporal attributes are not considered in computing dispersion. Toward that we use the Gini coefficient (Ceriani and Verme 2012); Gini co-efficient is convenient since it yields a normalized dispersion value wherein 0 implies perfect equality (minimal dispersion), and 1 indicates maximal inequality (high dispersion). The Gini index computation can be extended to multi-dimensional parameter vectors (Gajdos and Weymark 2005), which makes it suitable for our purpose. For our experiments, in which each location had a single parameter value, we use the following formula for calculating the Gini index (Deaton 1997)

$$gini(X_1, \dots, X_N) = \frac{N+1}{N-1} - \frac{2}{N(N-1)u} (\sum_{i=1}^N P_i X_i)$$

where N is number of data points in the cluster, u is the mean of the distribution and P_i is the rank of the data point after sorting the points in the population.

4.2 Anomaly detection

Once the clusters are formed, the second stage of our algorithm involves identifying which of these clusters are in fact anomalous. Toward that, we employ the LRT statistic (Mood et al. 1963). LRT is a standard significance test⁸ used to compare two nested models (or in this case, two distributions for their similarity) and is represented by D as:

$$D = -2 \ln \left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} \right)$$

Following the procedure explained in Pang et al. (2011), we assume that each cluster has an underlying Poisson distribution $P(\lambda_r)$, where λ_r is derived from the mean of the parameter values present in the cluster. A similar distribution is defined for the neighborhood region as $P(\lambda_n)$. We then compute the log-likelihood ratio, testing whether λ_r and λ_n are similar (null hypothesis) or differ significantly (alternative hypothesis). The test statistic value is then compared against the χ^2 value corresponding to a desired statistical significance (Huelsenbeck and Crandall 1997).

For each cluster identified through the earlier step, we compute the neighborhood region from the input grid by choosing appropriate *width* ρ as explained in Sect. 3. We

⁸ http://en.wikipedia.org/wiki/Statistical_model#Model_comparison.

then use LRT to figure out whether the distributions across the two samples are similar using the strategy explained earlier. If the LRT statistic value is above a threshold, γ , we identify the cluster as an anomaly (Line 17), and the final list of anomalies is then output in Line 18.

4.3 Discussion and analysis

4.3.1 Complexity

A brute-force method to figure out the optimal clustering over arbitrary shapes in a grid would be exponential in the size of the grid (Duczmal 2004). Our greedy strategy grows clusters by expanding into the neighbors based on a homogeneity condition. Let m be the number of neighbors for any grid cell; a cluster consisting of p cells would then have at most $p * m$ neighbors to expand into. At any iteration, there are $p * m$ Gini-index computations to find the closest neighbor, each computation being in $O(p \log p)$. The number of iterations is bounded by n , since each iteration accounts for exactly one cell. Therefore, the overall complexity of our approach is roughly $O(n m p^2 \log p)$. Clearly, if the grid is partitioned into extremely small clusters (i.e., small p), our algorithm would run with quasilinear complexity. Though the number of neighbors m is exponential in the number of spatio-temporal dimensions considered, in real-world application scenarios, it would be a fairly small number.

4.3.2 Thresholds

The threshold value chosen for Gini-based clustering can impact the cluster formation. More formally, the clusters generated using a higher Gini threshold are expected to be larger than those obtained using smaller thresholds. This is so since more heterogeneity can be tolerated under a larger threshold, and consequently the stopping condition is reached much later than with the case of a smaller threshold. Similarly, the threshold (z -value) chosen for LRT arbitrates labeling of a cluster as an anomaly. Choosing a higher threshold for LRT would result in fewer anomalies.

5 Experimental evaluation

In this section, we present the experimental evaluation of our approach for determining *spatio-temporal* anomalies over two real-world datasets.

5.1 Experimental setup

For empirical evaluation, we used two datasets. The first is a Climate Dataset⁹ which represents the entire globe divided in a 720×360 grid. We refer to this as *Dataset*₁. The cell values represent the temperature for a given temporal snapshot. For *spatial*

⁹ http://climate.geog.udel.edu/~climate/html_pages/download.html#ghcn_T_P2.

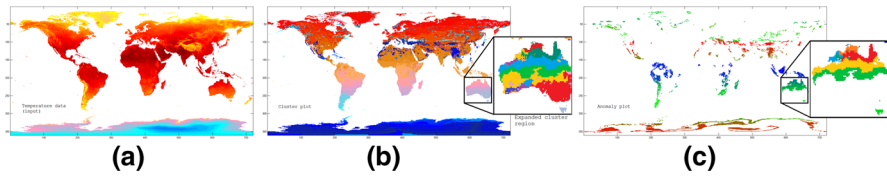


Fig. 7 **a** June 1981 grid snapshot, **b** phase one clusters, **c** phase two anomalies

anomaly detection, we select a grid representing a specific temporal instance (e.g., Fig. 7a for June 1981). For *spatio-temporal* anomalies, we select the grid values for the month of June over 12 years (1981–1992). Since no temperatures are reported for oceans, for each grid cell, only terrestrial neighbors were considered.

The second dataset pertains to ocean-bed topography in the region of Indian Ocean¹⁰. We refer to this dataset as *Dataset₂*. This dataset is a shelf bathymetry for the Indian Ocean region (20° E to 112° E, 38° S to 32° N) and is derived by digitizing the depth contours and sounding depths less than 200 m from the hydrographic charts published by the National Hydrographic Office, India. The depths are recorded at 5 *arcminute* intervals, resulting in a 1104 × 840 sized grid. The data generation details are described in [Sindhu et al. \(2007\)](#). This is a single snapshot dataset, and we used it for our spatial experiments.

Since the Gini co-efficient that we use requires non-negative values, we add an offset to all temperature/depth readings in these datasets such that all readings become non-negative. Unless mentioned otherwise, we use a value of 0.01 for the Gini indexing threshold τ and 3.84 for the LRT threshold; this LRT threshold corresponds to a statistical significance of 95 %.

5.1.1 User study

Apart from illustrative examples showing the working of the anomaly detection techniques, we also report user study results in our experimental evaluation. We conducted two user studies, both of which were directed at eliciting information from humans on the anomalousness of the anomalies identified by the different approaches. We created a web-survey for the study, and circulated it among the employees of our organization (i.e., IBM India Research Lab) through a broadcast email. In each of the two survey questionnaires, users were presented with a visual representation of the anomalies and asked to rate the quality of each anomaly on a 10-point scale. At the interest of keeping the instructions simple, we just asked the users to quantify the anomalous nature by comparing the candidate anomaly with its neighborhood. In particular, we did not inform the participants about the *generalized neighborhood* and hence, users could legitimately even rate transitional regions (e.g., Fig. 4a) as anomalies. We do not have the identities of the users who took the survey; however, the survey audience (i.e., to whom the email was sent) were mostly researchers with either a masters or doctoral degree in computer science or electrical engineering.

¹⁰ http://www.nio.org/index/option/com_subcategory/task/show/title/Sea-floorData/tid/2/sid/18/thid/113.

Table 2 Comparison with baselines on *Dataset₁* and *Dataset₂*

| Technique | <i>Dataset₁</i> | | <i>Dataset₂</i> | |
|---------------|----------------------------|--------|----------------------------|--------|
| | Mean | Median | Mean | Median |
| Our method | 5.94 | 6.21 | 8.05 | 8 |
| Local SaTScan | 3.00 | 3.00 | 2.85 | 3 |
| HAC-A | 1.66 | 1.47 | 2.25 | 2 |
| HC | 2.46 | 2.29 | 5.14 | 5.5 |

5.2 Spatial anomaly detection

The output at the end of each (of the two) phases is shown to illustrate the working of our algorithm. Figure 7b shows the homogeneous regions (*i.e.*, clusters) discovered at the end of the cluster formation phase. Unlike Fig. 7a, no specific color-coding scheme is employed apart from ensuring that adjacent clusters are assigned different colors. Since the total number of clusters is extremely large, two unrelated clusters may be represented by a single color. Figure 7c shows the filtered list of clusters at the end of the second phase, and represent the final list of anomalies that satisfy the LRT threshold. Please note that the colors are not indicative of the actual temperature, but similar color over a contiguous region indicates a cluster. However, similar color over two disjoint regions indicates two separate clusters independent of each other.

5.2.1 Comparison with baselines

We evaluated our approach against four different approaches¹¹ on both the datasets: (a) Local SaTScan, (b) HAC-A, (c) homogeneous clusters (HC) and (d) Image Segmentation. Local SaTScan is identical to the approach described in Kulldorff (1997) except that we apply LRT test to compare the sampled circular region against the generalized local neighborhood (defined in Sect. 3) instead of a global neighborhood. HAC-A, (outlined in Sect. 2), is the HAC variant that restricts pairwise mergers to only adjacent clusters. The output clusters are then ranked using a sum of size and $(1 - gini)$ where *gini* denotes the gini index within the cluster; this intuitively favors large and homogeneous clusters. HC represents phase one of our approach where the output clusters are ranked, by favoring large and homogeneous clusters. For Image Segmentation, we used a region detection technique (Achanta et al. 2009). Unlike other approaches, the input and output are both images; thus, instead of comparing ranked list of anomalies, we limit our comparison to a visual analysis of the output.

We conducted a user study among five users to compare the results of our approach against the baselines. We collected the top-7 anomalies from each technique (*i.e.*, Local SaTScan, HAC-A, HC and ours), and asked users to rate them on a 10 point scale (1 indicating *definitely not* anomalous, and 10 being perfect anomaly). Table 2 shows the results of our comparison; our technique is seen to achieve a mean score of

¹¹ We do not include outlier detection techniques in our comparative analysis since it is not clear as to how outlier detection techniques that estimate divergent behavior at each data object level may be fairly compared with techniques that discover groups of objects that exhibit divergent behavior.

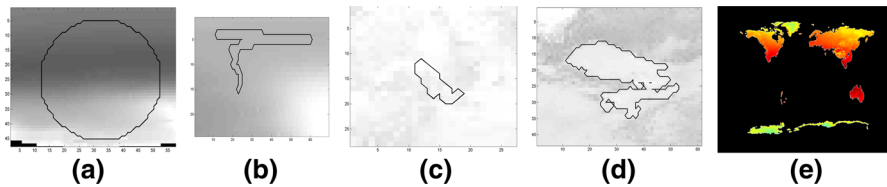


Fig. 8 The top-1 anomaly on *Dataset₁* from **a** local SaTScan, **b** HAC-A, **c** HC, **d** proposed method and **e** output from Blob detection

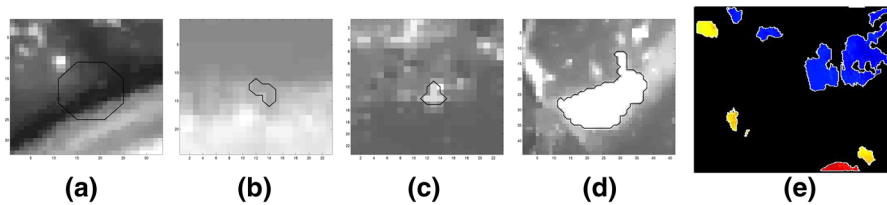


Fig. 9 The top-1 anomaly on *Dataset₂* from **a** local SaTScan, **b** HAC-A, **c** HC, **d** proposed method and **e** output from Blob detection

6 in the first dataset and 8 in the second (as much as twice the score of the second best technique). The top anomaly from the Local SaTScan, HAC-A, HC and our methods over *Dataset₁* are shown in Fig. 8a–d respectively. Figure 8e illustrates the image segmentation results, where each large colored component represents a single region. It may be judged that the results are unimpressive as they hardly seem to be anomalous regions, with large continents (*e.g.*, the entire North America, and North-central Asia) being put together into a single region. Thus, our analyses are seen to confirm that our technique is able to detect anomalous regions better than existing ones. The top anomaly from the Local SaTScan, HAC-A, HC and our methods over *Dataset₂* are shown in Fig. 9a–d respectively. Figure 9e illustrates the image segmentation results using the same blob detection technique.

5.2.2 Quality study

In addition to the above analysis, we performed a larger study with 36 users for *Dataset₁*¹². Given the shortcomings of using just phase one (as seen by the relatively poor ratings for HC in Table 2), we intended to use this study to evaluate the accuracy of (the LRT test for the) second phase of our approach. We took the ranked output from our technique, and selected 1–7, 41–47 and 81–87 ranked anomalies (total 21 candidates); we will refer to these as top (G_1), average (G_2), and low (G_3) ranked anomalies. The participants were requested to rank based on the degree of anomalousness (with 1 and 10 signifying *not an anomaly* and *perfect anomaly* respectively), as in the previous study; the results are summarized in Table 3.

¹² It must be noted that conducting user surveys is a difficult task. Hence, we conducted the user survey on *Dataset₁* only and not on *Dataset₂*.

Table 3 Quality assessment of anomalies by group with t test statistic value for significance of the results on *Dataset₁*

| Group | Anomaly ranks | Average score | t test stat |
|-------|---------------|---------------|--------------------|
| G_1 | 1–7 | 5.925 | 0.003 (vs. G_2) |
| G_2 | 41–47 | 5.218 | 0.195 (vs. G_3) |
| G_3 | 81–87 | 4.922 | – |

Table 4 The number of users who agreed upon a particular ranking for each group of anomalies on *Dataset₁*

| Group | High | Medium | Low |
|-------|------|--------|-----|
| G_1 | 24 | 6 | 6 |
| G_2 | 5 | 21 | 10 |
| G_3 | 7 | 9 | 20 |

It can be seen that G_1 (top-7 anomalies) received the highest mean score. To verify whether the results were significant, we analyzed them using the t test statistic¹³. Lower values of the t test statistic are desirable since they indicate that the scores being better due to chance are lower; the last column in Table 3 lists the values of the t test statistic illustrating that the better scores achieved by G_1 over G_2 are statistically significant too. Furthermore, although the scores for G_2 anomalies do not appear to be *significantly* better than the scores for G_3 anomalies, they are at least as good as latter. This confirms that the LRT test was able to rank anomalies in sync with the user perception. The highest average score for an individual anomaly (rank 3 from G_1) is approximately 8.5. This indicates that our approach not only ranks the anomalies appropriately, but it also detects significant anomalies.

Additionally, to assess the reliability of agreement among the surveyors, we calculated the Fleiss' kappa coefficient (κ) (Joseph 1971). For every user, we calculated the average scores for G_1 , G_2 and G_3 to categorize them in a relative ranking of *high*, *medium* and *low*. For example, if the average score of G_1 is better than G_2 and G_3 , and that of G_2 is better than G_3 , it implies that G_1 has *high*, G_2 has *medium*, and G_3 has *low* ranking. For every group, we quantified the number of users who agreed upon each of these rankings. The resultant matrix is shown in Table 4, and the κ value, bounded by 1 in case of complete inter-annotator agreement, evaluates to 0.149, which translates to slight inter annotator-agreement (Landis and Koch 1977). Further, we excluded the ratings of six annotators who largely contradicted the overall ratings, since these could be erroneous or due to a misunderstanding of the kind of anomalies we were looking for. After removing these, the κ value evaluates to 0.421, translating to a moderate agreement.

5.2.3 Scalability study

In order to assess the scalability of our technique, we analyzed the runtimes of our method. We varied the data size (i.e., the number of grid cells) from 1,000 to 1 million, by taking parts of, or piecing together consecutive snapshots of the climate dataset to form a squarish grid. The runtimes are tabulated in Table 5 for varying levels of

¹³ http://en.wikipedia.org/wiki/Students_t-test.

Table 5 Scalability tests: time in seconds

| Data size | Gini threshold | | |
|-------------------|----------------|-------|--------|
| | 0.0001 | 0.001 | 0.01 |
| 1,000 (1 k) | 0.215 | 0.590 | 1.256 |
| 5,000 (5 k) | 1.132 | 1.696 | 7.997 |
| 10,000 (10 k) | 1.710 | 3.107 | 12.845 |
| 1,00,000 (100 k) | 7.896 | 17.24 | 98.131 |
| 10,00,000 (1 mln) | 293.6 | 425.1 | 1392.3 |

Gini thresholds (for the first phase). The approach is seen to take in the order of a few seconds to a few minutes. It may be noted that the runtimes are not very critical since anomaly detection is expected to be an offline task to filter regions to feed to human agents who may want to analyze them further.

5.2.4 Stability study

In the first phase of our approach, i.e., cluster formation, we choose cluster seeds in no particular order. Thus, it is presumable that a different choice of cluster seeds could lead to a different clustering of cells at the end of the first phase. However, what we are more concerned about, is the stability of the top anomalies (i.e., output from the second phase) with respect to variations in the choice of seeds. Towards analyzing this, we made a row-major ordering of the cells in the grid, and called *Collections.shuffle*¹⁴ ten separate times, leading to ten different orderings. We then ran the technique ten times by using each of the ten orderings separately, and collected the list of the top- k anomalies from each. In particular, in the first phase, after each cluster is formed, the next cell from the input ordering that is yet unclustered is used as the seed for the next cluster. If the technique were completely insensitive to the ordering, any two runs would have one-to-one correspondence between the anomalies at any rank. However, in realistic scenarios, we do not expect a perfect match, but, expect that a top-ranked anomaly (we use top-ranked to mean ranked within k) from one run has a good match with a top-ranked anomaly from the other run. In particular, for a given value of k , we quantify this notion as follows:

$$val(i, j, k) = \frac{1}{k} \sum_{x=1}^k \max\{J(x^{th} \text{ anomaly from run } i, y^{th} \text{ anomaly from run } j) | 1 \leq y \leq k\}$$

where $J(., .)$ denotes the Jaccard similarity between the anomalies supplied to it. Informally, the above computation pairs each of the top-ranked anomalies from the i^{th} run with the best matching one from among the top-ranked anomalies in the j^{th} run. Then, the average of the similarity of the top-ranked anomalies of the i^{th} run with

¹⁴ [http://docs.oracle.com/javase/6/docs/api/java/util/Collections.html#shuffle\(java.util.List\)](http://docs.oracle.com/javase/6/docs/api/java/util/Collections.html#shuffle(java.util.List)).

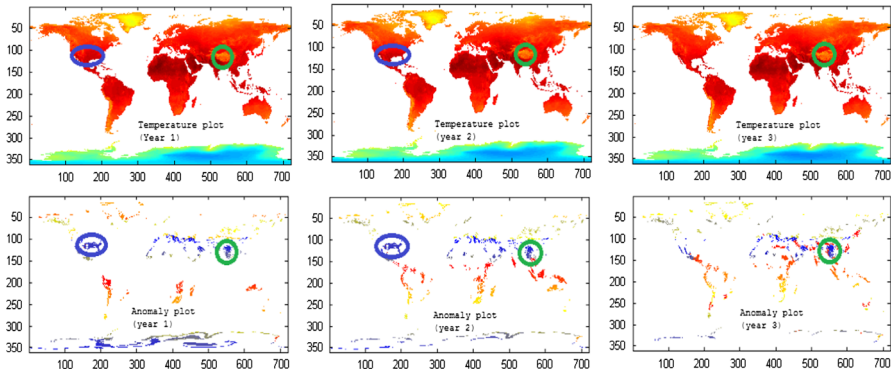


Fig. 10 Spatio-temporal anomalies: over three successive snapshots on *Dataset₁*

its paired anomaly (from the j th run) is computed. For each value of k , we aggregate $val(., ., k)$ over all the 90 pairs by simply averaging them:

$$aggrval(k) = \frac{1}{90} \sum_{1 \leq i \leq 10} \sum_{1 \leq j \leq 10, i \neq j} val(i, j, k)$$

Thus, $aggrval(k)$ computes the average match between an anomaly in the top- k of a run with its best matching pair in the top- k of another run, where the runs use different orderings of cells. Figure 11 plots the trends of $aggrval(k)$ against varying values of k from 1 to 20. For very low values of k , it is less likely that an anomaly from one run can find a good match in the other one (since only very few anomalies are considered); however, even at $k = 1$ when only the best anomaly is considered, it is seen that a high average overlap is recorded between the various runs ($aggrval(1) = 0.93$). This is seen to improve upto 0.98 at $k = 3$ beyond which the correlation between runs starts to decline. This could be due to the fact that as k increases, the anomalies are not that distinctive and there is more probability of being replaced by some other anomaly in the top- k list leading to a lower score. Given the very high overlap between the top anomalies, our technique may be considered to be stable with respect to choices of seeds.

5.3 Spatio-temporal anomaly detection

As outlined in Sect. 4, our approach is generalizable to the temporal dimension. Toward that, we selected, from *Dataset₁*, the monthly snapshot of June over a range of 12 years (*viz.*, 1981–1992). This enables meaningful comparison across years without being hampered by seasonal temperature variations. We presume it is much easier to visualize and understand a spatio-temporal anomaly when it is represented as a spatial anomaly that spans for a given time interval, rather than one that shrinks and grows in space with varying time. Accordingly, we constrain the cluster formation in the first phase so that temporal expansions always expand the whole cluster; for example, in the first temporal expansion of a spatial rectangular cluster, expansion on the temporal

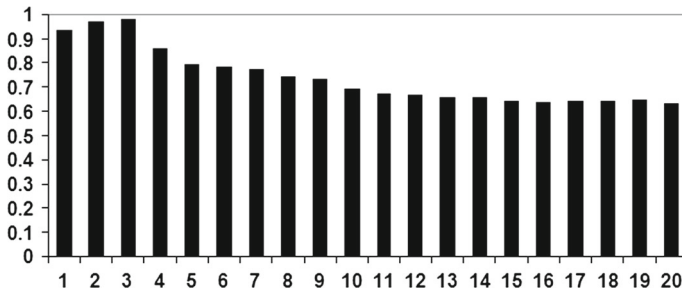


Fig. 11 Stability study: $aggrval(k)$ (Y-axis) versus k (X-axis)

dimension is constrained so that the cross-section of the temporally extended cluster is the rectangle itself. This also means that we would only discover spatio-temporal anomalies that have not moved in space with the passage of time; identification of anomalies that have grown/shrunk/move with time is not addressed in this work.

Figure 10 shows the anomalies obtained across three snapshots under this setting. The top row shows temporally consecutive snapshots of the data, whereas the bottom shows the spatio-temporal anomalies. We highlight two large anomalies among the top-ranked ones; blue ovals in the North American region which persist in the first two snapshots, and green circles in South-West China persisting across all the three snapshots. The corresponding regions in the top row are also highlighted with similar colors. The contrasting nature of these regions with their respective neighborhoods corroborates our results.

6 Conclusion & future work

In this paper we presented an automated domain-independent method for detecting *homogeneous* spatio-temporal anomalies that differ in behavior from their *local generalized neighbors*. In contrast to existing works that analyze *spatial* and *temporal* anomalies in isolation, we focused on detecting *spatio-temporal* anomalies within a single setting. Toward that, we proposed a two-step approach involving clustering and statistical dispersion and divergence tests. The experimental evaluation reveals that our approach performs better than existing state-of-the-art approaches.

We would like to point out that there are a few limitations of our method. For growing the clusters temporally, we require that consecutive snapshots have similar values. However, it may happen that a region is anomalous spatially over a time period, but have different values over time. For example, there may be a region that is always hotter compared to its neighbors, but the actual temperature varies over time. Our existing method will not be able to extend such an anomaly over the time period on which it holds. In our experiments, we got around this problem by considering periodic snapshots that correspond to similar time periods (*e.g.*, snapshot for the month of June for a series of years). There are two possible ways to address this problem and extend our method to a general temporal setting involving snapshots from a contiguous period (*e.g.*, every month of a year). One approach is to normalize the grid values at

each snapshot, so that they become comparable across time. An alternative strategy could be to inspect each snapshot independently, and then merge the anomalies across neighboring snapshots if they exhibit similar deviation from their spatial neighborhood and have the same shape across the snapshots. In this strategy, the values are never compared across time, only the shape of the anomaly is compared. We will explore these alternatives as part of the future work.

Furthermore, our experimentation was performed primarily on gridded weather data. Analyzing the generic nature of the problem and the applicability of our proposal to different domains renders as an interesting piece of work for further study. Specifically, detecting spatio-temporal anomalies in the context of traffic congestion monitoring, cellular network data analysis, disease outbreak detection and other such problem scenarios seems an interesting thread for future work. Further, developing efficient algorithms for detecting spatio-temporal anomalies in real time is an interesting problem to address.

References

- Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. 2012 IEEE Conf Comput Vis Pattern Recognit 0:1597–1604
- Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell 33(5):898–916
- Birant D, Kut A (2007) St-dbscan: An algorithm for clustering spatial-temporal data. Data Knowl Eng 60(1):208–221
- Bonett DG (2006) Confidence interval for a coefficient of quartile variation. Comput Stat Data Anal 50(11):2953–2957
- Bonnet N, Cutrona J, Herbin M (2002) A no-threshold histogram-based image segmentation method. Pattern Recognit 35(10):2319–2322
- Ceriani L, Verme P (2012) The origins of the gini index: extracts from *variabilità e mutabilità* (1912) by Corrado gini. J Econ Inequal 10(3):421–443
- Cheng T, Li Z (2004) A hybrid approach to detect spatial-temporal outliers. In Proceedings of the 12th International Conference on Geoinformatics Geospatial Information Research, pp. 173–178.
- Deaton A (1997) The analysis of household surveys: a microeconomic approach to development policy. Johns Hopkins University Press, Baltimore
- Duczmal L (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Comput Stat Data Anal 45(2):269–286
- El-Hamdouchi A, Willett P (1989) Comparison of hierarchie agglomerative clustering methods for document retrieval. Comput J 32(3):220–227
- Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, pp. 226–231
- Fan J, Yau DK, Elmagarmid AK, Aref WG (2001) Automatic image segmentation by integrating color-edge extraction and seeded region growing. IEEE Trans Image Process 10(10):1454–1466
- Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. Int J Comput Vis 59(2):167–181
- Friedman JH, Fisher NI (1999) Bump hunting in high-dimensional data. Stat Comput 9(2):123–143
- Gajdos T, Weymark JA (2005) Multidimensional generalized Gini indices. Econ Theory 26(3):471–496
- Grady L, Schwartz EL (2006) Isoperimetric graph partitioning for image segmentation. IEEE Trans Pattern Anal Mach Intell 28(3):469–475
- Huelsensbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. Ann Rev Ecol Syst 28:437–466
- Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
- Joseph FL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
- Kisilevich S, Mansmann F, Nanni M, Rinzivillo S (2010) Spatio-temporal clustering: a survey. Data mining and knowledge discovery handbook. Springer, New York, pp 855–874

- Kou Y, tien Lu C (2006) Spatial weighted outlier detection. In Proceedings of SIAM Conference on Data Mining
- Kulldorff M (1997) A spatial scan statistic. *Commun Stat-Theory Methods* 26(6):1481–1496
- Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(11):159–174
- Lukasová A (1979) Hierarchical agglomerative clustering procedure. *Pattern Recognit* 11(5–6):365–381
- Mankiewicz R (2000) The story of mathematics. Princeton University Department of Art, Princeton
- Mood A, Graybill F, Boes D (1963) Introduction to the theory of statistics. Mc-graw hill book company. Inc., New York
- Neill DB, Moore AW (2004) Rapid detection of significant spatial clusters. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pp. 256–265, New York, NY. ACM.
- Neill DB, Moore AW, Cooper GF (2005) A bayesian spatial scan statistic. In NIPS
- Ohlander R, Price K, Reddy DR (1978) Picture segmentation using a recursive region splitting method. *Comput Gr Image Process* 8(3):313–333
- Pang LX, Chawla S, Liu W, Zheng Y (2011) On mining anomalous patterns in road traffic streams. In *Advanced Data Mining and Applications*, pp. 237–251. Springer
- Patil GP, Taillie C (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environ Ecol Stat* 11:183–197
- Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: explorations in urban data collection. *IEEE Pervasive Comput* 6(3):30–38
- Revol C, Jourlin M (1997) A new minimum variance region growing algorithm for image segmentation. *Pattern Recognit Lett* 18(3):249–258
- Schubert E, Zimek A, Kriegel H-P (2014) Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Min Know Discov* 28(1):190–237
- Shekar S, Lu C-T, Zhang P (2002) Detecting graph-based spatial outliers. *Intell Data Anal* 6(5):451–468
- Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905
- Sindhu B, Suresh I, Unnikrishnan A, Bhatkar N, Neetu S, Michael G (2007) Improved bathymetric datasets for the shallow water regions in the indian ocean. *J Earth Syst Sci* 116(3):261–274
- Stolorz PE, Nakamura H, Mesrobian E, Muntz RR, Shek EC, Santos JR, Yi J, Ng KW, Chien S-Y, Mechoso CR, Farrara JD (1995) Fast spatio-temporal data mining of large geophysical datasets. In KDD, pp. 300–305
- Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. *Int J Health Geogr* 4:11