Contextual Spatial Outlier Detection with Metric Learning

Guanjie Zheng
College of Information Sciences and Technology
Pennsylvania State University
gjz5038@ist.psu.edu

Susan L. Brantley
Department of Geosciences
Pennsylvania State University
sxb7@psu.edu

Thomas Lauvaux
Department of Meteorology and Atmospheric Science
Pennsylvania State University
tul5@psu.edu

Zhenhui Li
College of Information Sciences and Technology
Pennsylvania State University
jessieli@ist.psu.edu

ABSTRACT

Hydraulic fracturing (or "fracking") is a revolutionary well stimulation technique for shale gas extraction, but has spawned controversy in environmental contamination. If methane from gas wells leaks extensively, this greenhouse gas can impact drinking water wells and enhance global warming. Our work is motivated by this heated debate on environmental issue and focuses on general data analytical techniques to detect anomalous spatial data samples (e.g., water samples related to potential leakages). Specifically, we propose a spatial outlier detection method based on contextual neighbors. Different from existing work, our approach utilizes both spatial attributes and non-spatial contextual attributes to define neighbors. We further use robust metric learning to combine different contextual attributes in order to find meaningful neighbors. Our technique can be applied to any spatial dataset. Extensive experimental results on five real-world datasets demonstrate the effectiveness of our approach. We also show some interesting case studies, including one case linking to leakage of a gas well.

KEYWORDS

Outlier detection; metric learning

1 INTRODUCTION

Improvements in high volume hydraulic fracturing, i.e., "fracking", which allows the development of shale gas, have changed the energy landscape. In 2000, only 1% of the natural gas production of U.S. is from shale gas, but this ratio reached over 20% by 2010 [37]. The U.S. government predicts that shale gas will make up 46% of U.S. natural gas production by 2035 [37]. However, fracking has spawned controversy about potential impacts on water quality and greenhouse gas emissions. Specifically, the most common water quality problem related to fracking in the biggest shale gas play (the Marcellus) is the escape of *methane* into surface and ground waters. If gas well leakage is large enough, it will impact individual aquifers, including homeowner wells in addition to impacting the rate of global warming [42].

Our work is motivated by this critical real-world environmental concern. Collaborating with geoscientists, we aim to detect anomalous water samples with high methane values. Such anomalies could potentially help us identify gas well leakage. While we have already published some preliminary results on searching for anomalous areas [29], in this paper we aim to pinpoint individual anomalous data samples. More generally, we tackle the problem of spatial outlier detection with contexts.

Specifically, the input of our outlier detection problem is a spatial dataset. For each data sample, we have a behavioral attribute (e.g., methane concentration in water), a sample location (i.e., GPS coordinates), and a set of additional contextual attributes describing each data sample (e.g., distance to the gas wells and nearby geological features). Our goal is to detect the anomalous data samples.

In the literature, typical outlier detection methods define the outliers as the samples that deviate significantly from the rest of the samples [3, 12]. Our problem is different because we target at the behavioral attribute (e.g., methane concentration in water samples) and are only interested in samples with unexpected high (or low) behavioral attribute values compared with a context (samples with similar contextual attributes). While we may directly identify samples with extremely high behavior attribute values (ignoring the contextual attributes), this will often provide trivial global outliers where the problems are already known (e.g., due to a known serious well leakage). Geoscientists are more interested in detecting non-trivial, local outliers for unknown leakage.

In order to detect anomalous behavioral attribute values with respect to a specific context, a simple baseline is to learn a regression model that predicts the behavioral attribute value using the contextual attributes as features. A data sample with observed value deviating significantly from its predicted value is then regarded as an outlier. However, in many real-world data, *the contextual features may not be informative enough to learn a reliable regression model.* For example, in our `Water` dataset, the methane concentration in groundwater is essentially unpredictable because many determining factors are either unknown (e.g., underground geology) or not well documented (e.g., anthropogenic activities like coal mining, industrial waste, and old residential houses). Hence, the outliers detected based on such a regression model may not be meaningful.

To overcome this difficulty, our intuition is to utilize the special property of spatial data – "near things are more related than distant things" [40]. We propose to use nearby data samples to identify

KDD '17, August 13–17, 2017, Halifax, NS, Canada.

DOI: http://dx.doi.org/10.1145/3097983.3098143

local outliers. While geo-coordinates have previously been used in the literature of spatial outlier detection to find spatial neighbors [35], *existing studies have not exploited the additional contextual features*. For example, two water samples might be spatially close but one is sampled from a swamp whereas the other one is sampled from a residential house. Intuitively, such additional contextual information can help us better identify neighbors in order to support outlier detection.

However, combining and utilizing all the contextual attributes (including spatial attributes) for neighborhood discovery is a non-trivial task, because the values of different contextual attributes carry different meanings and have different scales. While existing studies in contextual outlier detection all use Euclidean distance to find neighbors [12, 30], we propose to use metric learning to learn the distance metric. As metric learning enables us to assign different weights to different attributes, the neighbors we find are more meaningful. *To the best of our knowledge, this is the first work to use metric learning for outlier detection.*

Meanwhile, to successfully detect outliers in a local neighborhood, we are facing two additional challenges. *First, since the dataset contains outliers, the distance metric learned by traditional metric learning methods may not be reliable.* To address this challenge, in this paper we propose a new robust metric learning method in the regression setting. *Second, current methods do not differentiate the types of neighborhoods.* Some neighborhoods are more consistent, giving us more confidence to declare an outlier. Meanwhile, some neighborhoods are highly heterogeneous and may not provide enough confidence for us to detect outliers. Therefore, we further propose a method which incorporates a confidence score for each neighborhood to detect outliers.

To verify the effectiveness of our proposed method, we conduct extensive experiments on five real-world datasets and compare it with nine existing outlier detection methods. Further, we show two interesting case studies. We have successfully detected an abnormal water sample with potential leakage problem, and we are planning field trips to investigate it.

In summary, the key contributions of this paper are:

- We propose a local neighborhood-based method which combines heterogeneous contextual attributes with learned distance metrics to detect outliers. To the best of our knowledge, we are the first to apply metric learning techniques to outlier detection.
- We show how to address two challenging issues in the local neighborhood-based outlier detection, namely, distance metric learning with outliers and varying confidence levels in neighborhoods.
- We conduct extensive experiments on real-world datasets to demonstrate the effectiveness of our method. Interesting case studies are reported, demonstrating the uses of our method in real world scenarios including helping address a critical environmental problem related to fast shale gas development.

The rest of the paper is organized as follows. Related studies are first discussed in Section 2. Section 3 presents our problem definitions. We describe our method in Section 4 and discuss the experimental results in detail in Section 5. Finally, we conclude the paper in Section 6.

2 RELATED WORK

2.1 Outlier detection

Typical outlier detection methods aim to find data samples that are significantly different from other samples [12]. Classical methods include local outlier factor method [10] and high dimensional outlier detection [4]. However, such outlier detection methods do not differentiate contextual attributes and behavioral attributes and thus the outlier definition is quite different from ours.

Contextual outlier detection. Our outlier definition is similar to the definition of contextual outlier in the literature [12, 30]. In such problem settings, each data sample has contextual attributes and behavior attributes. A typical method is to first fit a predictive model using the contextual attributes as features and the behavioral attributes as response [23, 36]. The outlier score of a data sample is then calculated based on the predictive model. However, in real world applications we may not be able to obtain a reliable model due to missing contextual attributes (i.e., unknown factors).

Another group of methods first find the neighbors for each sample based on contextual attributes and then generate an estimation of the behavioral attribute using the neighbors [30, 41]. Outlier score is obtained by comparing the observed behavioral attribute values with the estimations. But these methods simply use Euclidean distance to combine different attributes. To handle heterogeneous attributes, in this paper we propose to use metric learning to detect more meaningful neighbors.

There are also works which define the contextual neighbors in graphs [20, 44] or multi-dimensional categorical data [39], but they are different from the numerical data used in our paper.

Spatial outlier detection. Spatial outliers are the data samples whose non-spatial attribute values are significantly different from their spatial neighbors [35]. Different methods are proposed to find spatial neighbors, e.g., kNN [14], Self-Organizing-Map [11], and graph-based method [27, 32]. Then, different statistic measures are applied to compare samples with their neighbors, e.g., Z-Score [5, 14], Mahalanobis distance [14], LOF-based measure [13, 24], and GLS-SOD [15]. However, all these methods only consider spatial information to find neighbors. We propose to further consider additional contextual attributes to find more precise neighbors.

2.2 Metric learning

Metric learning is a useful technique to learn a meaningful distance measure based on the data. The methods can be divided to unsupervised metric learning (e.g., Principle Component Analysis) and supervised metric learning. We only focus on the latter, which is more relevant to our problem. Typical metric learning methods [8, 28, 46] take similar pairs and dissimilar pairs as input, and learn a distance metric to make similar pairs closer to each other and make dissimilar pairs further apart. [45] extends metric learning to kNN kernel regression and learn the metric by minimizing kNN regression error. Our work is the first to apply metric learning to outlier detection problem.

Methods have been further proposed for robust metric learning. But current studies all assume pairwise similarity information is given [31, 43], or require discrete labels [19, 22]. Instead, we propose a robust metric learning technique for kernel regression.

3 PROBLEM DEFINITION

Suppose that we have a spatial data set of n data points $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$. Each data point $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ is composed of a contextual attribute vector $\mathbf{x}_i \in \mathbb{R}^d$ (including spatial coordinates) and a behavioral attribute value $y_i \in \mathbb{R}$. We will also use the notation $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ and $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ to represent the contextual attribute vector set and behavioral attribute value set correspondingly. For simplicity, we only focus on the case that there is just one behavioral attribute. But our work can be easily extended to multiple behavioral attributes. We summarize the notations in Table 1. We use bold lowercase letter $\mathbf{a}$ to represent a vector and bold uppercase letter $\mathbf{A}$ to represent a matrix. For better readability, we omit the transpose notation in writing $\mathbf{z}_i = (\mathbf{x}_i, y_i)$.

Table 1: Notation used for problem definition and method description.

$\mathbf{x}_i$	Contextual attribute vector of sample i
y_i	Behavioral attribute value of sample i
$\hat{y}_i$	Estimation of behavioral attribute value of sample i
$\mathbf{z}_i = (\mathbf{x}_i, y_i)$	Data point for sample i
C_i	Local confidence in determining the outlier score of sample i
S_i	Outlier score of sample i
N_i	The set of k-nearest contextual neighbors of sample i
w_{ij}	Weight for sample j in estimating the behavioral attribute value of sample i
d_{ij}	Distance between sample i and sample j
$\mathbf{M}$	Learned distance metric
$\mathbf{A}$	Distance metric projection matrix

Our problem can now be formulated as follows:

Problem 1 (Contextual Outlier Detection). *Given data set $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$, we wish to assign an outlier score $S_i \in [0, 1]$ to each data point $\mathbf{z}_i$, using $\mathbf{x}_i$ as contextual attributes and y_i as the behavioral attribute. Higher score indicates that this sample has a higher probability to be an outlier.*

In the context of `Water` dataset of methane measurements (refer to Section 5 for detailed data description), for a data sample $\mathbf{z}_i$, contextual attributes $\mathbf{x}_i$ can include latitude, longitude, and distance from the sample location to shale gas wells. Behavioral attribute y_i is the methane concentration measured from water samples.

In the context of a real estate dataset (refer to `Zillow` dataset in Section 5), contextual attributes $\mathbf{x}_i$ describe a real estate's properties such as latitude, longitude, square feet, and year of built, and behavioral attribute y_i can be the sold price. We will use the `Zillow` dataset as the illustrating example throughout the paper to explain our method, because this dataset is easier to understand compared with the `Water` dataset, which contains many geoscience terms.

4 CONTEXTUAL OUTLIER DETECTION

Given the corresponding contextual attributes, the most straightforward way to determine the abnormality of one behavioral attribute value is to learn a regression model $f : \mathbb{R}^d \to \mathbb{R}$ that predicts the behavioral attribute value using the contextual attributes as features: $\hat{y}_i = f(\mathbf{x}_i)$. Then, the outlier score can be defined as the difference between the true response value y_i and the prediction $\hat{y}_i$. However, this approach is subject to at least two major issues in practice. First, the underlying relationship among the contextual and behavioral attributes may be highly complicated, and there is no evidence that such relationship can be captured by a global regression model. Second, some indicative features could be missing in real world datasets. For example, in `Zillow` dataset, we do not have a quantitative measure on the interior decoration of the house or the competitiveness of the local real estate market, thus it is hard to predict an accurate price. Similarly in the `Water` dataset, factors such as underground geology and anthropogenic activities are unknown and thus it is hard to predict the methane concentrations in groundwater.

4.1 Local Models

To address the aforementioned challenges, we make two key observations about the data. *First, while it is often difficult to build a global regression model, samples in a local neighborhood of the data space may be well approximated by a local regression model. Second, since we are focusing on spatial datasets, we can utilize the spatial neighbors with similar contextual attributes to help us build a more accurate model.* According to the First Law of Geography by Waldo Tobler [40], "everything is related to everything else, but near things are more related than distant things". Hence, we can assume that data samples that are spatially close to each other should share similar properties, even though some properties are not observed. For example, in the `Zillow` dataset, it is reasonable to assume that houses in same community have similar house properties. In the `Water` dataset, groundwater samples collected at geographically close locations should share similar underground geology.

The above observations motivate us to develop local models which first find contextual neighbors for each data sample, and then use the behavioral attributes of these neighbors to predict the behavioral attribute value for that particular data sample. Specifically, we use the kNN kernel regression [45]:

$$\hat{y}_i = \frac{\sum_{j \in N_i} w_{ij} y_j}{\sum_{j \in N_i} w_{ij}}, \tag{1}$$

where w_{ij} is the Gaussian kernel weight

$$w_{ij} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) \tag{2}$$

defined according to some distance measure d_{ij} between any two samples, and σ is the standard deviation of the distance distribution. In addition, N_i denotes the set of k-nearest contextual neighbors of sample $\mathbf{z}_i$:

$$N_i = \{j : \mathbf{z}_j \text{ is among the samples with the } k \text{ smallest } d_{ij}\}. \tag{3}$$

4.2 Robust Metric Learning for Contextual Neighborhood Discovery

It is easy to see that our local method heavily depends on the distance measure d_{ij} for contextual neighborhood discovery. In

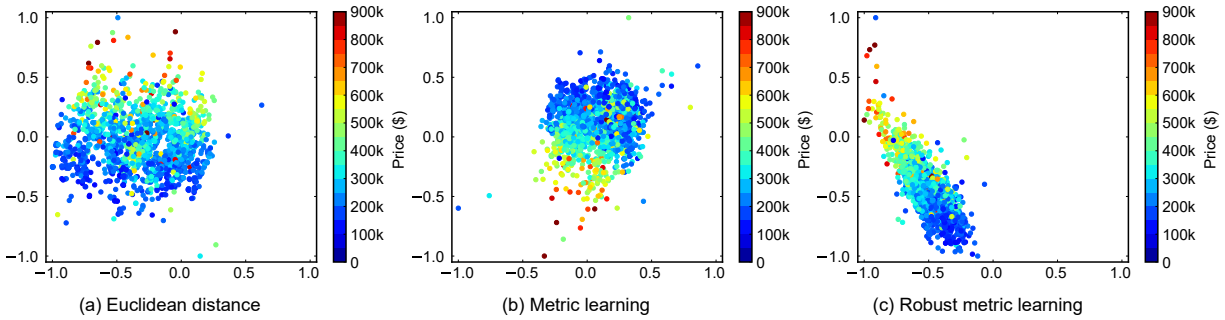


Figure 1: Illustration of distance calculation on `Zillow` dataset using different distance metrics. The data points are projected onto 2D dimension using MDS (Multidimensional Scaling). The color of the points represents the sold price of the houses.

practice, Euclidean distance metric is commonly used to combine different features and compute d_{ij}. However, there are several important drawbacks of this standard metric. Specifically, in real world scenarios, it is often the case that there are features with totally different semantic meanings in a dataset. For example, in `Zillow` dataset, geo-location (latitude and longitude), year of built, and square feet all stand for very different aspects of the properties of the house. When calculating the distance between two houses, for example, it is non-trivial to combine spatial distance and the time difference in year of built. Further, there may be features which are not particularly relevant to the regression task, but the Euclidean distance metric would assign them with the same weight as other features.

Therefore, we propose to learn a more meaningful distance measure using metric learning. The intuition is to adjust the weight on each dimension of features when calculating the distance between samples, so that similar data points (in terms of label in classification or response in regression) will become closer and dissimilar samples will be pushed further away from each other.

Mathematically, the Mahalanobis distance between two data points is defined as (following similar notations as in [45]):

$$d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}}^2 = \|(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)\|, \tag{4}$$

where $\mathbf{M}$ can be any symmetric positive semi-definite matrix. The matrix $\mathbf{M}$ can be further decomposed into the product of two matrices [45]:

$$\mathbf{M} = \mathbf{A}^T \mathbf{A}, \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{d \times r}$ can also be regarded as a projection matrix that projects the original d-dimensional space onto a r-dimensional space. Hence, we have

$$d_{ij}^2 = \|(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A}^T \mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\| = \|\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\|^2. \tag{6}$$

Note that, by setting $\mathbf{A} = \mathbf{I}_d$, this distance metric degenerates to Euclidean distance.

Therefore, the goal of metric learning for kernel regression [45] is to learn the distance metric $\mathbf{M}$ (or equivalently projection matrix $\mathbf{A}$), so that the regression error can be minimized. Typically, the quadratic regression loss is commonly used [45]:

$$\mathcal{L} = \sum_i (y_i - \hat{y}_i)^2. \tag{7}$$

Given the loss function, the projection matrix $\mathbf{A}$ can be learned by a gradient decent algorithm [45]:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 4\mathbf{A} \sum_i (\hat{y}_i - y_i) \sum_j w_{ij} (\hat{y}_i - y_j)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \tag{8}$$

4.2.1 Robust Metric Learning. In the literature, the projection matrix $\mathbf{A}$ is learned with the assumption that the data samples are free from corruptions. However, this is not the case for our outlier detection problem. For example, in the `Zillow` dataset, the sold price of some properties are actually the price of the real estate lot, instead of the house. Therefore, their prices might look much lower than expected. In `Water` dataset, some groundwater samples may be polluted by shale gas leakage and hence have a very high methane concentration, compared to nearby samples. As existing metric learning methods also include these samples, the learned matrix $\mathbf{A}$ could be significantly biased.

However, it is well known that the quadratic loss Eq. (7) is sensitive to outliers [9]. Therefore, we replace it with ℓ_1-norm loss and use the following loss function to approximate ℓ_1-norm loss in order to make it convenient to do derivative calculation:

$$\mathcal{L} = \sum_i \sqrt{(y_i - \hat{y}_i)^2 + \xi}, \tag{9}$$

where ξ is a small constant making the objective function differentiable. In our experiments we set ξ to 0.0001. The value of ξ does not influence the experimental results significantly.

Correspondingly, the new gradient can be derived as follows:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 4\mathbf{A} \sum_i (\hat{y}_i - y_i) \sum_j \frac{\hat{y}_i - y_j}{2\sqrt{(y_i - \hat{y}_i)^2 + \xi}} w_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \tag{10}$$

Example 4.1. We use the `Zillow` dataset to illustrate the effect of the proposed robust metric learning. We inject 2% outliers into the dataset (details about outlier injection can be found in Section 5.3.1). The distribution of samples under different distance metrics are shown in Figure 1. We observe that, in Figure 1(a), under Euclidean distance, houses with high sold prices and low sold prices are mixed together. Moving from Figure 1(a), (b) to (c), i.e., from Euclidean distance, metric learning, to robust metric learning, houses with different sold prices are more separated from each other, and houses

with similar sold prices are closer to each other. Therefore, the distance metric between samples is better modeled.

Finally, we note that while we focus on contextual outlier detection in this paper, the proposed robust metric learning approach is quite general and can be applied to many outlier detection methods, such as distance based methods [7, 26], density-based methods [10], kNN-based methods [30]. To the best of our knowledge, previous outlier detection methods have never used metric learning to learn the distance between samples. We are the first to apply metric learning techniques to outlier detection.

4.3 Outlier Score with Local Confidence

After obtaining the behavioral attribute estimation $\hat{y}_i$ for each data sample, the outlier score can be simply defined as the difference between $\hat{y}_i$ and the groudtruth behavioral attribute y_i:

$$S_i^G = |y_i - \hat{y}_i|. \tag{11}$$

However, in spatial datasets, such a definition ignores the heterogeneity inside regions. For example, in `Zillow` dataset, some regions might be a mixture of houses in different styles and levels, which makes the price in this region highly unpredictable (i.e., many houses have high S^G). In this case, even if we find a house whose true price deviates a lot from its estimation, we are not very sure if this is a true outlier. Meanwhile, if we find such a house in a region where the prices of most houses can be well estimated (i.e., S^G is low for most houses in that region), we have a higher confidence that this house is an outlier.

Therefore, we further define the local confidence of the data sample:

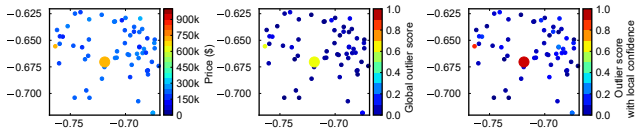
$$C_i = \frac{1}{\sum_{j \in N_i} S_j^G}. \tag{12}$$

Then, we define the outlier score with local confidence as follows:

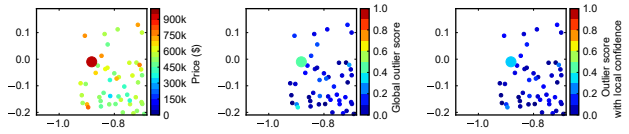
$$S_i^L = S_i^G \times C_i = \frac{S_i^G}{\sum_{j \in N_i} S_j^G}. \tag{13}$$

Example 4.2. We use real examples in `Zillow` dataset to illustrate the effect of considering the local confidence for defining the outlier score. We show two houses in Figure 2. House 1 has a much higher sold price ($715k) compared to its neighborhood (most are lower than $300k). In addition, the global outlier score S^G of House 1 is relatively high (0.63). Considering the fact that all the houses in the neighborhood has a global outlier score lower than 0.1, we have more confidence to assign this house a high outlier score (0.93).

House 2 also has a very high sold price ($949k) compared to its neighborhood. However, the house price in its neighborhood varies from $15k to $90k, which indicates the houses in this region are very heterogeneous. Consequently, houses in this region have global outlier scores varying from 0 to 0.5. Therefore, although House 2 has a global outlier score 0.43, we have little confidence to detect it as an outlier. Considering the local confidence, we assign it a lower outlier score 0.32.



(a) House 1 (price: $715k, global outlier score: 0.63, outlier score with local confidence: 0.93)



(b) House 2 (price: $949k, global outlier score: 0.45, outlier score with local confidence: 0.32)

Figure 2: Illustration of defining outlier score with local confidence on `Zillow` dataset. The data points are projected onto 2D dimension using MDS. The color of the points represents the sold price, global outlier score S_i^G, and outlier score with local confidence S_i^L of the houses, respectively. The big dot in the middle is the target house.

5 EXPERIMENT

In this section, we conduct experiments on five real datasets. We show a comprehensive quantitative evaluation by comparing with other methods and also show some interesting case studies. [1]

5.1 Datasets

All the five real datasets contain spatial information (i.e., latitude and longitude) as part of the contextual attributes. The `Water` dataset and `Air` dataset are provided by our collaborators in geoscience and meteorology. The datasets report the methane concentrations in groundwater and in atmosphere and allow exploration of the impacts of shale gas development. `Zillow` dataset, which contains online real estate information, is collected from Zillow API [25]. The other two datasets, `El Nino` and `Hydro`, are obtained from UCI repository [2].

In order to construct the features for `Water` dataset and `Air` dataset, we define the notation as in Table 2.

Table 2: Notation for feature construction in `Water` and `Air` datasets.

e_j	Emission volume of source j
N_i^e	Neighboring emission sources around sample i
w_{ij}^e	Weight for emission source j to sample i
$\vec{d}_{ij}^e$	Distance vector from emission source j to sample i
$\vec{u}_i$	Wind vector at sample i

`Water` dataset [1]. This dataset contains 1,645 data samples of methane concentration in groundwater in Pennsylvania. Detailed description of this dataset can be found in [29]. According to the geologists, among all the factors that might lead to abnormal methane

[1] Code and data are available at the authors' website.

concentration in the groundwater, gas wells (including conventional gas wells and unconventional gas wells) and faults are believed to be the most important ones. Therefore, for each groundwater sample, we construct 11 contextual attributes (i.e., features) describing sampling location (*latitude* and *longitude*) and nearby emission sources, including *distances to nearest gas wells*, *density of gas wells* (Eq. (14)), *emission intensity of gas wells* (Eq. (15)), *number of gas wells in certain distance threshold*, *total emission volume of gas wells in certain distance threshold*, and *distances to nearest faults*. The behavioral attribute is the methane concentration measured in groundwater sample. The outliers found in this dataset could potentially indicate shale gas leakage problem.

$$\text{density of gas wells around sample } i = \sum_{j \in N_i^e} w_{ij}^e, \tag{14}$$

$$\text{emission intensity of gas wells around sample } i = \sum_{j \in N_i^e} w_{ij}^e e_j, \tag{15}$$

where $w_{ij}^e = g(\|\vec{d}_{ij}^e\|)$ is the Gaussian weight with ($\mu = 0, \sigma = 5$km):

$$g(d) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(d-\mu)^2}{\sigma^2}\right). \tag{16}$$

`Air` dataset [38]. This dataset contains 34,100 data samples of the methane concentration in the atmosphere collected in the New York State and Pennsylvania. The detailed description of this dataset can be found in [6]. According to the geologists, in addition to the many natural emission sources, which are usually not well documented, conventional gas wells, unconventional gas wells, industrial emissions and gas compressors are the major documented emission sources that may contribute to the methane concentration change in the air. Meanwhile, the methane concentration is strongly correlated to the meteorological features. Therefore, we use 30 contextual attributes in this dataset, including *sampling location* (*latitude* and *longitude*), *sampling time*, *meteorological features*, *geographical features*, *distances to emission sources*, and *source emission intensity considering wind speed and direction* (Eq. (17)). Similar to `Water` dataset, the behavioral attribute is methane concentration measured in atmosphere.

$$\text{emission intensity around sample } i \text{ considering wind} = \sum_{j \in N_i^e} w_{ij}^e e_j \tag{17}$$

with

$$w_{ij}^e = g(\|\vec{d}_{ij}^e\|) \cdot \cos\theta_{ij}. \tag{18}$$

Here, $g(\|\vec{d}_{ij}^e\|)$ is the Gaussian weight with ($\mu = 0, \sigma = \gamma \times \|\vec{u}_i\|$km), γ is a constant, and θ_{ij} is the angle between $\vec{u}_i$ and $\vec{d}_{ij}^e$.

`Zillow` dataset [25]. This dataset contains 1,511 house selling records in State College, Pennsylvania, from year 2014 to 2016. The sold price ranges from $100,000 to $975,000. The contextual attributes describing the real estate properties include latitude, longitude, square feet, year of built (7 attributes in total). We use the most recent sold price as the behavioral attribute.

`El Nino` dataset [17]. This dataset contains 93,935 samples in the equatorial Pacific. The contextual attributes are oceanographic and surface meteorological variables (6 contextual attributes in total) and the behavioral attribute is sea surface temperature.

`Hydro` dataset [34] contains 308 records describing the relationship between the shape of a ship and the residuary resistance that the ship bares in water. The longitudinal position of the buoyancy and five shape parameters of the ship are used as contextual attributes. The behavioral attribute is the residuary resistance.

5.2 Methods for Comparison

We compare the performance of our method, named MELODY (MEtric Learning Outlier Detection), with following four categories of outlier detection methods.

General outlier detection methods.

LOF [10] is one of the most frequently used outlier detection methods. It finds local outliers by comparing the local reachability density of each sample with its neighbors.

Contextual outlier detection methods.

CAD [36]: Conditional anomaly detection proposes to model the data using Gaussian Mixture Models (GMM). A GMM U is used to model contextual attributes, with U_i representing the i-th component. Another GMM V is used to model behavioral attributes, with V_j representing the j-th component. Then, a mapping function $p(V_j|U_i)$ is learned to give the probability that the behavioral attribute of a sample is generated by V_j given its contextual attributes are generated by U_i. The lower the probability that a sample is generated by this model, the higher its outlier score is.

ROCOD [30]: Robust contextual outlier detection proposes to simultaneously consider local and global effects in outlier detection. Specifically, kNN regression is used to generate a local estimation for each sample, and a ridge regression (ROCOD.RIDGE) or tree regression (ROCOD.CART) is used to produce a global estimation for each sample. These two estimations are then combined to generate a total estimation $\hat{y}_i$ for the behavioral attribute value. The outlier score is defined as $|y_i - \hat{y}_i|$.

Regression-based outlier detection methods.

LR: Linear regression. We use the contextual attributes as features and the behavioral attribute as the response variable. Then, the outlier score is defined as the absolute difference between the groundtruth and the estimated response variable values: $|y_i - \hat{y}_i|$.

XGBOOST [16]: XGBOOST is a gradient tree boosting method which achieves impressive accuracy in many classification and regression tasks in practice. We apply the same setting as in LR to learn the model. Parameters are selected based on cross-validation.

Spatial outlier detection methods. All these methods first use spatial attributes to find neighbors for each sample. The difference is that how they define outliers using neighbors.

ZS [35]: The spatial statistic $S(x)$ is defined as the difference between the behavioral attribute value y at location x and the average behavioral attribute value of its neighbors. The outlier score for the attribute is then defined as $Z_s(x) = \frac{S(x)-\mu_S}{\sigma_S}$, where μ_S and σ_S denote the mean and standard deviation of $S(x)$, respectively.

SOD [14] uses all the other attributes (except spatial attributes) as behavioral attributes. The behavioral attribute values for each sample are estimated by taking the mean or median of its neighbors'

Table 3: Overall performance comparison in terms of AUC on all datasets. The best performance of the compared methods is highlighted. Percentage in parenthesis is the relative improvement over the performance of the best baseline method.

Methods \ Datasets	Zillow	Water	Air	El Nino	Hydro
LOF	0.159	0.071	0.024	0.616	0.422
CAD	0.354	0.110	**0.244**	0.439	0.146
ROCOD.CART	0.422	0.121	0.208	0.595	0.769
ROCOD.RIDGE	0.403	0.118	0.104	0.333	0.611
LR	0.389	0.114	0.057	0.285	0.770
XGBOOST	**0.477**	0.119	0.083	**0.780**	**0.935**
ZS	0.206	0.122	0.219	0.622	0.234
SOD	0.167	0.054	0.034	0.292	0.487
GLS-SOD	0.188	**0.142**	0.208	0.619	0.254
MELODY	**0.687 (+44%)**	**0.182 (+28%)**	**0.716 (+193%)**	**0.970 (+24%)**	**0.965 (+3.2%)**

attribute values. The outlier score is defined as the Mahalanobis distance between observed and estimated behavioral attribute values.

GLS-SOD [15] uses the same behavioral attribute as our method. A local generalized least square regression model is used to model the behavioral attribute value variation over the space. The outlier scores are determined according to the standard estimated residuals.

5.3 Quantitative Evaluation

5.3.1 Experiment Setting. We use generated outliers in this section for two reasons. First, it is extremely hard to find spatial datasets with ground truth outliers. Second, typical outlier datasets usually contain global outliers, which is different from the task of detecting contextual outlier in this paper. Therefore, we follow [30, 36] to generate outliers using a perturbation scheme. Specifically, the original dataset (without perturbation) is assumed to contain no outlier. To inject one outlier, we first randomly select a sample $\mathbf{z}_i = (\mathbf{x}_i, y_i)$. Then, we randomly pick k_s (set to 10 by default) samples from the dataset, and select the sample $\mathbf{z}_j = (\mathbf{x}_j, y_j)$ with the maximum y value difference $|y_i - y_j|$ from these k_s samples. Then we replace $\mathbf{z}_i$ with $\mathbf{z}'_i = (\mathbf{x}_i, y_j)$ in the dataset.

By default, the ratio of injected outliers p is set to 2% for `Zillow`, `Water`, `Air` and `El Nino` datasets. For the `Hydro` dataset, we set $p = 5\%$ due to the smaller dataset size. The number of contextual neighbors k_N in our method is fixed to 60. For all the experiments, the results shown are the average of 20 runs of perturbation.

5.3.2 Evaluation Metrics. All the outlier methods output a full list of samples ranked by their outlier scores in descending order. We use Precision at K, Recall at K, and the Area Under the Curve (AUC) of the Precision Recall Curve (PRC) as the evaluation metrics. We use PRC instead of Receiver Operating Characteristic (ROC) because PRC provides a more informative evaluation of performance of algorithms in skewed datasets [18]. Note that in our experiments, the outlier samples only constitute a small portion of the dataset.

5.3.3 Overall Performance. We first compare MELODY with other methods on `Zillow` dataset. We show the Precision at K and Recall at K in Figure 3. The maximum K is set to 30 because the number of outliers is 30 in `Zillow` dataset. We observe that MELODY achieves higher precision and recall than all other methods.

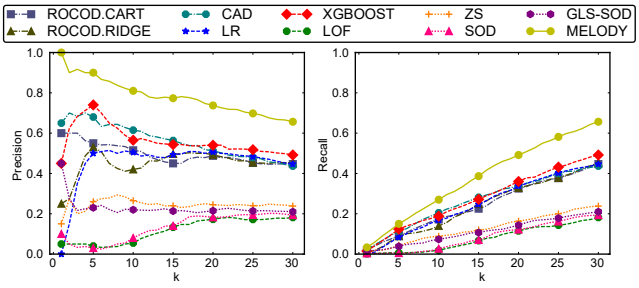


Figure 3: Performance comparison on `Zillow` dataset

Next, we show AUC values for all the methods on all datasets in Table 3. As one can see, MELODY consistently performs the best on all datasets. We further make several interesting observations on the results. First, all the methods perform poorly (AUC < 0.2) on `Water` dataset. To understand why this is the case, we show the R^2 for XGBOOST and MELODY in Table 4. The results suggest that `Water` dataset is indeed hard to model. Second, all other methods perform poorly on `Air` dataset (with highest AUC as 0.244), whereas MELODY is still very robust (with AUC as 0.776). We can also see in Table 4 that the regression fitting of MELODY is much better than XGBOOST on `Air` dataset. This indicates that it is necessary to use neighbor-based method for regression and outlier detection.

Table 4: Regression fitting results (R^2).

Method	Zillow	Water	Air	El Nino	Hydro
XGBOOST	0.690	0.027	-0.087	0.777	0.915
MELODY	0.715	0.091	0.646	0.947	0.955

5.3.4 Performance w.r.t. Outlier Ratio p. We next investigate the performance of all methods w.r.t. the number of injected outliers. As shown in Table 5, MELODY always performs the best among all methods regardless of the outlier ratios. We also observe that, with higher outlier ratios, the performance of all the methods are getting better in terms of AUC. This is because, with a higher outlier ratio,

the top candidates detected by the methods are more likely to be true outliers. In other words, all the methods will achieve higher precision with the same recall when more outliers are injected into the dataset. In the extreme case, if we set all the samples in the dataset to be outliers, the precision will always be 1.0 for all recall values, therefore the AUC value will also be 1.0.

Table 5: Performance w.r.t. outlier ratio p on `Zillow` dataset. Results shown are AUC on `Zillow` dataset.

Method	p=1%	p=2%	p=3%	p=4%	p=5%
LOF	0.123	0.159	0.196	0.219	0.245
CAD	0.310	0.354	0.368	0.376	0.406
ROCOD.CART	0.304	0.422	0.514	0.552	0.616
ROCOD.RIDGE	0.298	0.403	0.501	0.544	0.602
LR	0.288	0.389	0.482	0.522	0.584
XGBOOST	**0.374**	**0.477**	**0.569**	**0.593**	**0.645**
ZS	0.162	0.206	0.282	0.323	0.374
SOD	0.132	0.166	0.202	0.208	0.231
GLS-SOD	0.152	0.187	0.254	0.295	0.346
MELODY	**0.611**	**0.687**	**0.703**	**0.750**	**0.771**

5.3.5 Performance w.r.t. Perturbation Sampling Size k_s. Table 6 shows the performance of all methods w.r.t. perturbation sampling size k_s. As one can see, MELODY consistently outperforms other methods. We also observe that, as k_s increases, the performance of all the methods is generally getting better. This is because, when k_s increases, more samples will be considered as candidates for perturbation. Since we select the sample with the most different behavioral attribute value, it is more likely to use an extreme value. Consequently, the perturbed sample is more likely to be an obvious outlier. In addition, we note that, our method MELODY is particularly robust when the outliers are not obvious (i.e., with small k_s values). As shown in Table 6, the gap between MELODY and other methods is much larger when k_s is smaller.

Table 6: Performance w.r.t. Perturbation Sample Size k_s. Results shown are AUC on `Zillow` dataset. Relative improvement over best baseline in parenthesis. Smaller k_s values mean less obvious outliers.

Methods	k_s=10	k_s=20	k_s=30	k_s=40	k_s=50
LOF	0.159	0.194	0.210	0.216	0.210
CAD	0.354	0.349	0.312	0.327	0.312
ROCOD.CART	0.422	0.580	0.648	0.709	0.745
ROCOD.RIDGE	0.403	0.556	0.634	0.686	0.718
LR	0.389	0.539	0.619	0.661	0.697
XGBOOST	**0.477**	**0.626**	**0.718**	**0.771**	**0.804**
ZS	0.207	0.345	0.420	0.488	0.521
SOD	0.167	0.219	0.250	0.263	0.290
GLS-SOD	0.187	0.317	0.403	0.468	0.501
MELODY	**0.687** (+44%)	**0.766** (+22%)	**0.824** (+15%)	**0.848** (+10%)	**0.866** (+7.7%)

5.3.6 Justification of Each Component in MELODY. We have four components in MELODY. In this experiment, we verify the effectiveness of each component. The four components include (1) the use of contextual neighbors to estimate the behavioral attribute (denoted as CN); (2) metric learning to assign weights on different contextual attributes (denoted as M); (3) robust metric learning (denoted as RM); and (4) local confidence factor for outlier detection (denoted as L). As shown in Table 7, improvement in outlier detection performance is indeed achieved by including each component in our method.

Table 7: Performance comparison of adding each component (CN: contextual neighbors, L: local confidence, M: metric learning, RM: robust metric learning). Results shown are AUC on `Zillow` dataset.

Components	CN	CN+M	CN+RM	CN+RM+L
AUC	0.413	0.469	0.493	0.687

5.4 Parameter Sensitivity

The only parameter in MELODY is the number of contextual neighbors, denoted as k_N. In this section, we show how the outlier detection performance is affected by this parameter. In Figure 4, we can observe that AUC first increases and then decreases slowly as k_N increases. In the range of [30, 100], the performance is relatively stable, hence MELODY is not very sensitive to the choice of k_N.

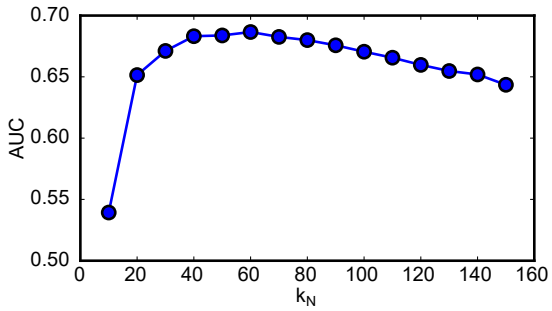


Figure 4: Performance w.r.t. number of neighbors k_N on `Zillow` dataset.

5.5 Case Study

In this section, we show two case studies on `Zillow` dataset and `Water` dataset, respectively. We use the original datasets without injecting any outliers. We show the cases with top outlier scores detected by MELODY. Maps are made from Google Maps [21].

5.5.1 Case study on Water Dataset. We show an interesting case on `Water` dataset in Figure 5. This water sample has a higher methane concentration compared to the "neighboring" samples. This is an outlier that the geologists believe could be related to methane leakage.

5.5.2 Case study on Zillow Dataset. Figure 6 shows a case study on `Zillow` dataset. This is a neighborhood on the east side of Penn State University in State College, PA. This house is similar to its neighbors, but has a much higher sold price.

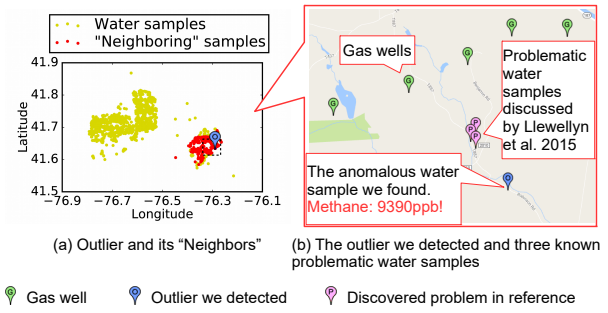


Figure 5: Case study on `Water` dataset. In Figure 5(a), the methane concentration of this sample ranks 66/1645 (top 4.0%) in the whole dataset (yellow dots and red dots), and ranks 8/300 (top 2.6%) among the "neighboring" samples (red dots). In Figure 5(b), we can observe that the detected outlier (blue balloon) is only 1km downstream from a site where we know that methane leaked into three homes (pink balloons) along a branch of Sugar Run in Terry township [33]. The methane concentration in the water of these sites is influenced by the upstream gas wells.

6 CONCLUSION

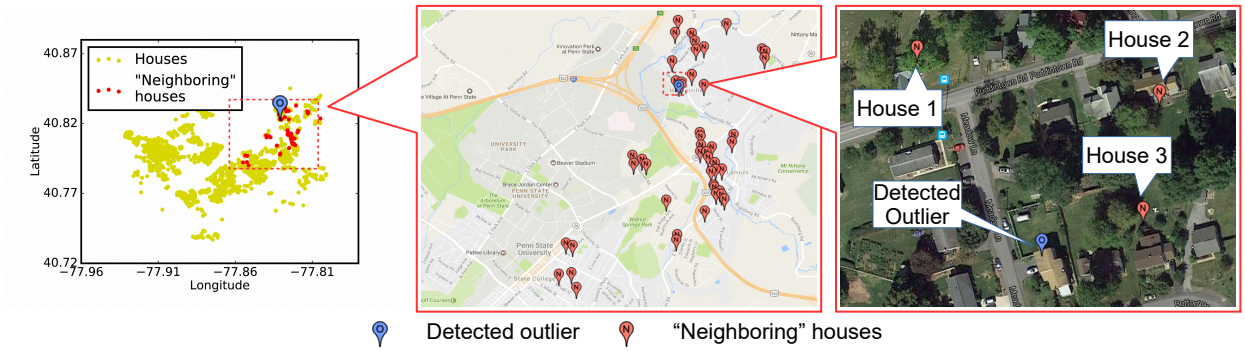
Our work is motivated by the environmental concern caused by shale gas development. We aim to detect outliers from spatial dataset. Different from existing outlier detection methods, we propose a local neighborhood-based method by combining heterogeneous contextual attributes via robust metric learning. Extensive experimental results demonstrate the effectiveness of our proposed method. The proposed technique is being used to help geoscientists locate potential environmental issues (e.g., gas well leakage).

ACKNOWLEDGMENTS

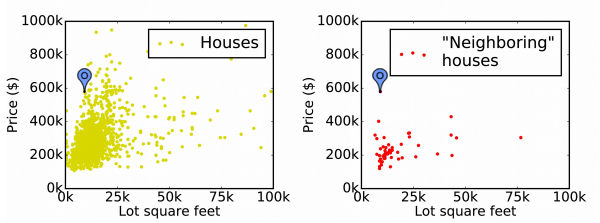
The work was funded from a gift to Penn State for the Pennsylvania State University General Electric Fund for the Center for Collaborative Research on Intelligent Natural Gas Supply Systems and was supported in part by NSF awards #1639150, #1618448, #1652525, and #1544455. This work has also been funded by the U.S. Department of Energy National Energy Technology Laboratory (project DE-744FE0013590). The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

REFERENCES

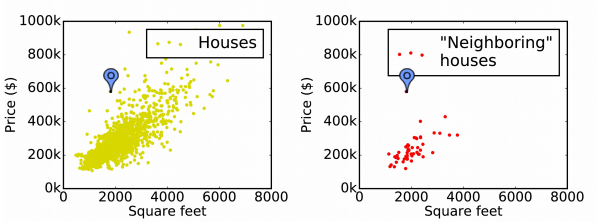
[1] 2015. Shale Network. (2015). https://doi.org/10.4211/his-data-shalenetwork
[2] 2017. UCI Machine Learning Repository. (2017). http://archive.ics.uci.edu/ml/
[3] Charu C Aggarwal. 2015. Outlier analysis. In *Data mining*. Springer, 237–263.
[4] Charu C Aggarwal and Philip S Yu. 2001. Outlier detection for high dimensional data. In *ACM Sigmod Record*, Vol. 30. ACM, 37–46.
[5] Luc Anselin, Ibnu Syabri, and Youngihn Kho. 2006. GeoDa: an introduction to spatial data analysis. *Geographical analysis* 38, 1 (2006), 5–22.
[6] Z. R. Barkley, T. Lauvaux, K. J. Davis, A. Deng, Y. Cao, C. Sweeney, D. Martins, N. L. Miles, S. J. Richardson, T. Murphy, G. Cervone, A. Karion, S. Schwietzke, M. Smith, E. A. Kort, and J. D. Maasakkers. 2017. Quantifying methane emissions from natural gas production in northeastern Pennsylvania. *Atmospheric Chemistry and Physics Discussions* 2017 (2017), 1–53. https://doi.org/10.5194/acp-2017-200
[7] Stephen D Bay and Mark Schwabacher. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 29–38.
[8] Aurélien Bellet, Amaury Habrard, and Marc Sebban. 2013. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709* (2013).
[9] Peter Bloomfield and William Steiger. 2012. *Least absolute deviations: Theory, applications and algorithms*. Vol. 6. Springer Science & Business Media.
[10] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *ACM sigmod record*, Vol. 29. ACM, 93–104.
[11] Qiao Cai, Haibo He, and Hong Man. 2013. Spatial outlier detection based on iterative self-organizing learning model. *Neurocomputing* 117 (2013), 161–172.
[12] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.
[13] Sanjay Chawla and Pei Sun. 2006. SLOM: a new measure for local spatial outliers. *Knowledge and Information Systems* 9, 4 (2006), 412–429.
[14] Dechang Chen, Chang-Tien Lu, Yufeng Kou, and Feng Chen. 2008. On detecting spatial outliers. *Geoinformatica* 12, 4 (2008), 455–475.
[15] Feng Chen, Chang-Tien Lu, and Arnold P Boedihardjo. 2010. Gls-sod: a generalized local statistical approach for spatial outlier detection. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1069–1078.
[16] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.
[17] Di Cook. 1999. UCI Machine Learning Repository. (1999). https://archive.ics.uci.edu/ml/datasets/El+Nino
[18] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 233–240.
[19] Zheyun Feng, Rong Jin, and Anil Jain. 2013. Large-scale image annotation by efficient and robust kernel metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 1609–1616.
[20] Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han. 2010. On community outliers and their efficient detection in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 813–822.
[21] Google. 2017. Google Maps. (2017). https://developers.google.com/maps/
[22] Ran He, Bao-Gang Hu, Wei-Shi Zheng, YanQing Guo, et al. 2010. Two-Stage Sparse Representation for Robust Recognition on Large-Scale Database.. In *AAAI*, Vol. 10. 1–1.
[23] Charmgil Hong and Milos Hauskrecht. 2015. MCODE: Multivariate Conditional Outlier Detection. *arXiv preprint arXiv:1505.04097* (2015).
[24] Tianqiang Huang and Xiaolin Qin. 2004. Detecting outliers in spatial database. In *Image and Graphics (ICIG'04), Third International Conference on*. IEEE, 556–559.
[25] Zillow Inc. 2016. Zillow. (2016). http://www.zillow.com/
[26] Edwin M Knorr, Raymond T Ng, and Vladimir Tucakov. 2000. Distance-based outliers: algorithms and applications. *The VLDB Journal—The International Journal on Very Large Data Bases* 8, 3-4 (2000), 237–253.
[27] Yufeng Kou, Chang-Tien Lu, and Raimundo F Dos Santos. 2007. Spatial outlier detection: a graph-based approach. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, Vol. 1. IEEE, 281–288.
[28] Brian Kulis et al. 2013. Metric learning: A survey. *Foundations and Trends® in Machine Learning* 5, 4 (2013), 287–364.
[29] Zhenhui Li, Cheng You, Matthew Gonzales, Anna K Wendt, Fei Wu, and Susan L Brantley. 2016. Searching for anomalous methane in shallow groundwater near shale gas wells. *Journal of Contaminant Hydrology* 195 (2016), 23–30.
[30] Jiongqian Liang and Srinivasan Parthasarathy. 2016. Robust Contextual Outlier Detection: Where Context Meets Sparsity. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2167–2172.
[31] Meizhu Liu and Baba C Vemuri. 2012. A robust and efficient doubly regularized metric learning approach. In *European Conference on Computer Vision*. Springer, 646–659.
[32] Xutong Liu, Chang-Tien Lu, and Feng Chen. 2010. Spatial outlier detection: Random walk based approaches. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 370–379.
[33] Garth T Llewellyn, Frank Dorman, JL Westland, D Yoxtheimer, Paul Grieve, Todd Sowers, E Humston-Fulmer, and Susan L Brantley. 2015. Evaluating a groundwater supply contamination incident attributed to Marcellus Shale gas development. *Proceedings of the National Academy of Sciences* 112, 20 (2015), 6325–6330.
[34] Roberto Lopez. 2013. UCI Machine Learning Repository. (2013). https://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics
[35] Shashi Shekhar, Michael R Evans, James M Kang, and Pradeep Mohan. 2011. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, 3 (2011), 193–214.



(a) The detected outlier and "Neighboring" houses on the map



(b) Sold price w.r.t. Lot square feet



(c) Sold price w.r.t. Square feet

House	Square feet	# Bedrooms	# Bathrooms	Lot square feet	Year built	Year sold	Price ($)	Zestimate ($)	Address
Detected outlier	1,801	5	2.5	9,148	1955	2016	580,000	234,056	120 Meadow Ln State College, PA 16801
House 1	1,778	3	2	8,712	1943	2016	120,000	192,430	1707 Puddintown Rd State College, PA 16801
House 2	1,840	4	2.5	11,326	1950	2014	181,000	222,883	1800 Puddintown Rd State College, PA 16801
House 3	2,818	5	3	11,761	1952	2014	215,000	329,950	310 Bottorf Dr State College, PA 16801

(d) Three houses that are very similar to the detected outlier

Figure 6: Case study on `Zillow` dataset. The detected outlier is similar to its neighbors in contextual attributes such as spatial location, "square feet" and "lot square feet", but its sold price is much higher. As shown in Figure 6(b), when we plot sold price v.s. lot square feet for the entire dataset, this house does not appear to be an outlier, indicating that this outlier can not be easily detected by global outlier detection methods (e.g., linear regression). By comparing it to the "neighboring" houses only, MELODY successfully detect this outlier. Similar observations can be made in Figure 6(c) w.r.t. square feet. Figure 6(d) shows that the three houses which are very close and similar to the detected outlier have much lower sold prices. In addition, the Zestimate values provided by Zillow also suggest that this is an outlier ($234,056 Zestimate vs. $580,000 sold price).

[36] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. 2007. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering* 19, 5 (2007), 631–645.

[37] Paul Stevens. 2012. The 'shale gas revolution': Developments and changes. *Chatham House* (2012), 2–3.

[38] Colm Sweeney, Anna Karion, Eric Kort, Mackenzie Smith, Tim Newberger, Stefan Schwietzke, Sonja Wolter, and Thomas Lauvaux. 2015. Aircraft Campaign Data over the Northeastern Marcellus Shale. (2015). https://doi.org/10.15138/G35K54

[39] Guanting Tang, Jian Pei, James Bailey, and Guozhu Dong. 2015. Mining multidimensional contextual outliers from categorical relational data. *Intelligent Data Analysis* 19, 5 (2015), 1171–1192.

[40] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.

[41] Michal Valko, Branislav Kveton, Hamed Valizadegan, Gregory F Cooper, and Milos Hauskrecht. 2011. Conditional anomaly detection with soft harmonic functions. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 735–743.

[42] Radisav D Vidic, Susan L Brantley, Julie M Vandenbossche, David Yoxtheimer, and Jorge D Abad. 2013. Impact of shale gas development on regional water quality. *Science* 340, 6134 (2013), 1235009.

[43] Hua Wang, Feiping Nie, Heng Huang, and H Huang. 2014. Robust Distance Metric Learning via Simultaneous L1-Norm Minimization and Maximization.. In *ICML*. 1836–1844.

[44] Xiang Wang and Ian Davidson. 2009. Discovering contexts and contextual outliers using random walks in graphs. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 1034–1039.

[45] Kilian Q Weinberger and Gerald Tesauro. 2007. Metric Learning for Kernel Regression.. In *AISTATS*. 612–619.

[46] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. 2002. Distance metric learning with application to clustering with side-information. In *NIPS*, Vol. 15. 12.