Outlier Detection from Non-Smooth Sensor Data

Timo Huuhtanen
Department of Computer Science
Aalto University
Espoo, Finland
firstname.lastname@aalto.fi

Henrik Ambos
Department of Computer Science
Aalto University
Espoo, Finland
firstname.lastname@aalto.fi

Alexander Jung
Department of Computer Science
Aalto University
Espoo, Finland
firstname.lastname@aalto.fi

***Abstract*—Outlier detection is usually based on smooth assumption of the data. Most existing approaches for outlier detection from spatial sensor data assume the data to be a smooth function of the location. Spatial discontinuities in the data, such as arising from shadows in photovoltaic (PV) systems, may cause outlier detection methods based on the spatial smoothness assumption to fail. In this paper, we propose novel approaches for outlier detection of non-smooth spatial data. The methods are evaluated by numerical experiments involving PV panel measurements as well as synthetic data.**

***Index Terms*—outlier detection, spatial signals**

I. Introduction

We consider datasets that are constituted by data points associated with particular spatial locations. Such datasets arise in various important applications including numerical weather prediction, economics, power grids and wireless sensor networks. Typically, spatial data is assumed to conform with the geometry of the spatial domain. Outlier detection from a dataset is an important data analysis task - to detect interesting events or hint at faulty devices. The basic idea underlying most outlier detection methods is to compare each data point with an expected or anticipated value due to some signal model. A large deviation of a data point from its expected value indicates that the data is an outlier. For spatial data, a widely used signal model is smoothness in the sense of requiring similar values for data points at close-by locations.

However, in some applications the generated data exhibits intrinsic discontinuities. For such non-smooth spatial data, existing methods relying on smoothness are ill-suited. We are using photovoltaic (PV) panel monitoring as an actual example case for spatial outlier detection. The measurement system is illustrated in Figure 1(a). Each PV panel is equipped with a sensor measuring the electrical power generated by the panel. As the location of each panel is different and known, the dataset produced by the sensors becomes a spatial dataset illustrated in Figure 1(b). The target of the monitoring is to identify malfunctioning PV panels - i.e. panels whose power is smaller than predicted. The problem can then be seen as an outlier detection from spatial dataset. The spatial dataset collected from the sensor network is usually smooth in spatial domain, because the adjacent PV panels are equally illuminated. However, shadows of nearby objects may cause part of the panels to be in shadow while others are in direct sunlight. Our research question is then, how can we construct a good outlier detector for a non-smooth spatial signal.

In this paper, we compare different approaches for the problem of spatial outlier detection in the presence of discontinuities. Our contributions in this paper are comparisons of different methods for this problem. According to our knowledge, this is the first attempt to treat spatial outlier detection as label propagation problem.

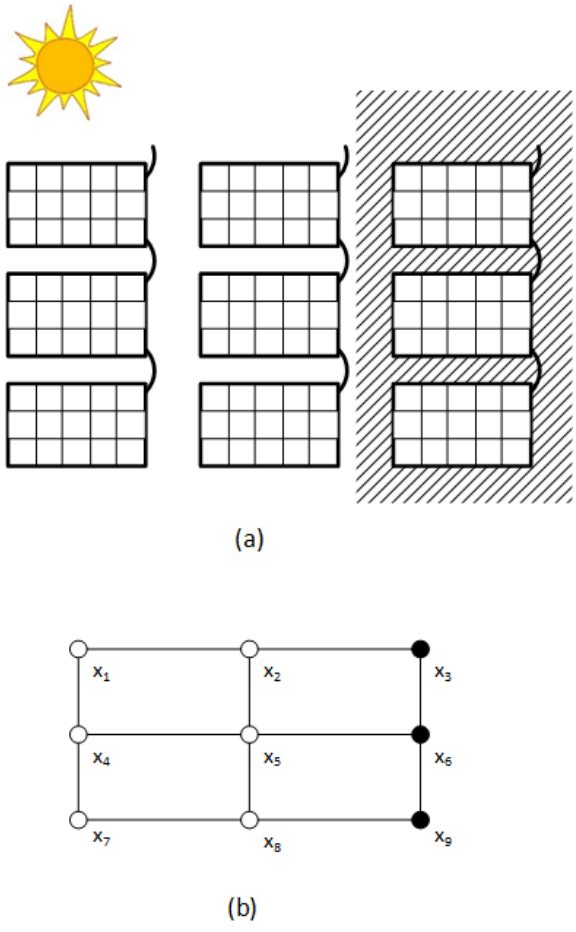

Fig. 1. PV panel monitoring as an example of outlier detection problem from a dataset: (a) PV panels, part of which are shadowed; (b) panel-level power measurements making a spatial dataset.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

II. Problem setting

Our data set consists of power measurement values of several PV panels. The panels have known spatial locations in a regular rectangular grid as illustrated in Figure 1(b). So, our data set consists of spatial data points (nodes) denoted as $\mathcal{V} = \{1, 2, ..., n\}$. The attribute x_i is a mapping from $\mathcal{V} \mapsto \mathbb{R}$ representing the attribute value of node i, in the PV monitoring case the power measurement at panel i.

Our target is to detect outlier nodes from the spatial data set. However, since our data set consists of normal data without outliers, we cannot use it directly for development of outlier detection algorithms. Instead, we use a common approach in outlier detection [1] and convert the outlier detection problem into prediction problem: we use the data points in the neighborhood of our target node, $\{j|j \in N(i)\}$, to construct a predictor for the attribute of our target node, denoted as $g(N(i))$. Since the actual measurement of the target node is known, we can compare the actual value, x_i to the predicted value, $g(N(i))$. A large difference between the predicted and actual measurement values indicates an outlier. As in [2], we can define an outlier score indicating the degree of outlierness, $o(i)$, for node i as

$$o(i) = |x_i - g(N(i))| \tag{1}$$

The spatial smoothness assumption is valid in the PV measurement data if all the nodes are in direct sunlight or in full shadow. However, the assumption fails in the partially shadowed situation illustrated in Figure 4. We express the shadowed/non-shadowed status of each node by assigning label, y_i, for each node: $y_i = -1$ indicates a shadowed node and $y_i = +1$ indicates a non-shadowed node.

The neighborhood relationships of the nodes can be represented with an undirected weighted data graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes of the graph and $\mathcal{E}$ is the set of edges between vertices. Here nodes are the data points and edges represent the similarity of data points. We define the edge weights via a thresholded Gaussian kernel weighting function as in [3],

$$W_{ij} = \begin{cases} e^{-\frac{(x_i - x_j)^2}{2\theta^2}} & \text{if } x_j \in N(i) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Here θ is a constant and $N(i)$ consists of four or eight closest neighbors of the data point as illustrated in Figure 2 and in Figure 3.

III. Proposed Methods

Our aim is to find a good predictor, $g(N(i))$, in the presence of edge-like spatial discontinuities in the data. We compare three different approaches. Our baseline approach is a maximum likelihood predictor, which is based on signal smoothness assumption the neighborhood and thus ignores the existence of discontinuities. Secondly, we apply methods from image processing based on explicitly detecting discontinuities, referred as edge detection, or implicitly taking them into account with median-based filter [4]. Thirdly, we try to solve the situation of shadowed/non-shadowed nodes over all the whole spatial area by using graph-based semi-supervised learning methods [5].

A. Maximum likelihood predictor

Spatial smoothness of the signal means that the signal is locally constant in the neighborhood, $N(i) \cup \{i\}$. We can model each x_i as an independent and identically distributed random variable, which is a sum of constant value, μ, and zero-mean Gaussian distributed noise, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. So, the probability density function for each x_j is,

$$f(x_j; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_j - \mu)^2}{2\sigma^2}} \tag{3}$$

In order to get an estimate for μ, we can construct the log-likelihood function as

$$log\mathcal{L}(\mu, \sigma; x) = log\Big(\prod_{j \in N(i)} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_j - \mu)^2}{2\sigma^2}} \Big) \tag{4}$$

Calculating $\frac{\partial}{\partial \mu} log\mathcal{L}(\mu, \sigma; x) = 0$ then gives the maximum likelihood estimate for μ, which is also the maximum likelihood predictor of x_i, which is the sample mean

$$\hat{\mu} = g_{ML}(N(i)) = \frac{1}{|N(i)|} \sum_{x_j \in N(i)} x_j \tag{5}$$

We can select $N(i)$ to be e.g. the 4-point spatial neighborhood illustrated in Figure 2 or the 8-point spatial neighborhood illustrated in Figure 3.

B. Maximum likelihood (ML) predictor with edge detection

If there is a discontinuity present inside the spatial neighborhood, as illustrated in Figure 4, it causes an abrupt change of the attribute value and the smoothness assumption is no longer valid. In order to be able to use the smoothness assumption we need to redefine the spatial neighborhood in such a way that only data points on the same side of the edge as x_i are included in the neighborhood. E.g. in the case of Figure 4 $N_{reduced}(i) = \{1, 2, 3\}$.

For detecting the edge, we use a simple thresholding based detector described in [6],

$$y_i = \begin{cases} +1, & x_i \geq thr \\ -1, & x_i < thr \end{cases} \tag{6}$$

and apply (5) into the reduced spatial neighborhood,

$$N_{reduced}(i) = \{j | j \in N(i) \wedge y_j = y_i\} \tag{7}$$

Several histogram-based methods for selecting the threshold are reviewed in [6]. Here we set the threshold halfway between minimum and maximum values found in the spatial dataset. From our prior knowledge of PV panels in partially shadowed conditions [7] we can conclude that this threshold lies between the power generated by shadowed and unshadowed panels.

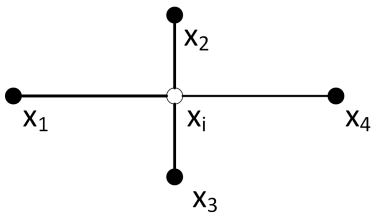

Fig. 2. 4-point spatial neighborhood of x_i.

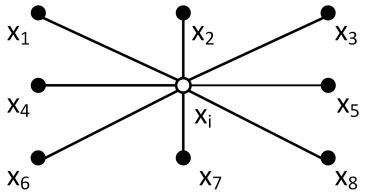

Fig. 3. 8-point spatial neighborhood of x_i.

C. Median based predictor

Median filtering has been widely used in image processing due to its capability to preserve edges in the image. On the other hand, sample median is known to be more robust than sample mean [8]. Here we try the following median-based predictor algorithm.

$$g_{med}(N(i)) = median(x_j | j \in N(i)) \tag{8}$$

D. Predictors based on semi-supervised learning of discontinuities

A drawback of the methods, which try to detect discontinuities based on the local spatial neighborhood of the data point, is that with noisy input they tend to give noisy detection results. Since the edge-like discontinuities are continuous in some spatial direction, a more reliable interpretation of the discontinuities in the sensor dataset can potentially be found by aiming at a global interpretation for the shadow structure of the data set. To this end, we use semi-supervised learning algorithms as follows:

1) We select randomly a set of data points and classify them with the thresholding method in (6).
2) We run a semi-supervised classification algorithm to find the labels for the rest of the data points. This algorithm will find the shadowed and unshadowed areas for the whole dataset. Thus, we can expect to get a global view into the shadow situation - which potentially is more reliable than the local methods.
3) We run predictor (5) with the reduced neighborhood (7) taking into account the global view into the shadow situation; i.e. only the data points of the neighborhood which are similarly shadowed as the target data point are included in the neighborhood.

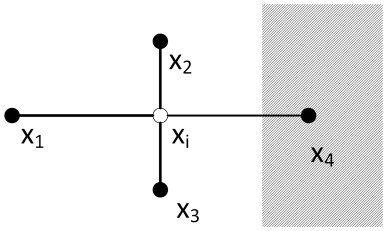

Fig. 4. 4-point spatial neighborhood of x_i with a discontinuity in the neighborhood.

We compare two different semi-supervised classification methods for this purpose: label propagation algorithm [9] and logistic network Lasso [10].

The known labels, y_i for the label propagation algorithm can be collected into the vector denoted as $Y = \{y_1, ..., y_L\}$. The estimated labels $\hat{Y} = (\hat{Y}_l, \hat{Y}_u)$ where $\hat{Y}_l$ includes the known labels and $\hat{Y}_u$ denotes the unknown labels. Label propagation uses affinity matrix, $\mathbf{W} = [W_{ij}]$, constructed from the edge weights (2) and diagonal degree matrix $\mathbf{D} = [D_{ii}]$, where

$$D_{ii} = \sum_j W_{ij} \tag{9}$$

The labels are found with the following iterative algorithm [9] which updates the labels in step 5, but forces the known labels back to the original values in step 6:

Algorithm 1 Label propagation

1: Compute affinity matrix $\mathbf{W}$ using (2)
2: Compute the diagonal degree matrix $\mathbf{D}$ using (9)
3: Initialize $\hat{Y}^{(0)} \leftarrow (y_1, ..., y_l, 0, ..., 0)$
4: **repeat**
5: $\quad \hat{Y}^{(t+1)} \leftarrow \mathbf{D}^{-1}\mathbf{W}\hat{Y}^{(t)}$
6: $\quad \hat{Y}_l^{(t+1)} \leftarrow Y_l$
7: **until** convergence to $\hat{Y}^{(\infty)}$
8: Label all the points x_i by the sign of $\hat{y}_i^{(\infty)}$

Logistic network Lasso [10] is a classifier which finds the values for the unknown labels by regularized empirical risk minimization

$$x = \arg\min_x (\hat{E}(x) + \lambda ||x||_{TV}) \tag{10}$$

where the empirical error is defined by

$$\hat{E}(x) = \frac{1}{|\hat{Y}_l|} \sum_{i \in \hat{Y}_l} log(1 + exp(-y_i x_i)) \tag{11}$$

and the total variation (TV) by

$$||x||_{TV} = \sum_{(i,j) \in \mathcal{E}} W_{ij} |x_i - x_j| \tag{12}$$

The labels are the achieved as $y_i = sign(x_i)$.

IV. Experiments

We evaluate the performance of the proposed algorithms using a synthetic data set and a real-world dataset obtained from PV monitoring system by SOLA Sense Ltd. We prepared two synthetic dataset consisting of 16-by-16 spatial samples including edge-like discontinuities in various directions: one with added Gaussian random noise ('edges+noise' illustrated in Figure 5) and another with added smooth slope ('edges+slope' illustrated in Figure 6).

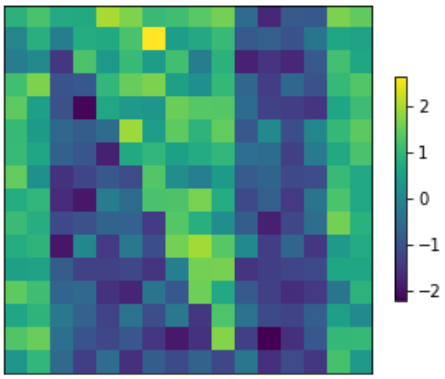

Fig. 5. A synthetic 16-by-16 point dataset ('edges+noise') used for experiments

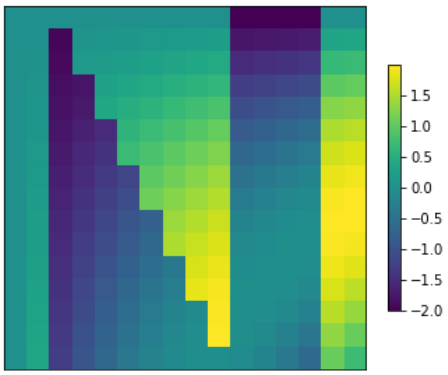

Fig. 6. A synthetic 16-by-16 point dataset ('edges+slope') used for experiments

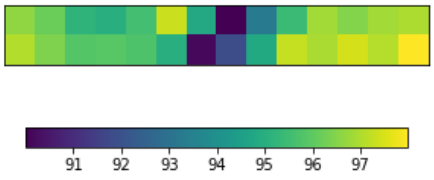

Fig. 7. One data frame from the PV data set used for experiments

The PV dataset includes power measurements of a 2-by-14 PV panel system. The dataset contains output power measurements from each panel at 10-second intervals over one day (6000 time instances corresponding to 17 hours of measurement). The selected day was sunny and the PV system has nearby objects causing regular shadows falling on the panels during the day.

Both datasets include only normal data without real outliers. However, we can use the average outlier score (1) calculated over the entire dataset to measure the goodness of our predictors (and thus also the outlier detectors based on the predictors). An ideal predictor predicts the actual attribute values perfectly giving an outlier score equal to zero. Thus, the smaller outlier score we achieve, the better our predictor is.

V. Results

The outlier scores of different algorithms for the synthetic test data sets ('edges+noise', 'edges+slope') as well as for the solar data set are listed in Table I. In order to get an idea of how the different predictors treat the edges in the spatial data, we plot the values of $g(N(i))$ for the different algorithms for data set 'edges+slope' in Figure 8.

TABLE I
Average outlier score of different prediction algorithms in 4-point spatial neighborhood(n4) and 8-point spatial neighborhood(n8)

Algorithm	***Synthetic edges+noise***	***Synthetic edges+slope***	***PV data***
Maximum likelihood(n4)	0.548	0.285	2.386
ML with edge detection(n4)	0.445	**0.025**	2.104
4-point median(n4)	0.515	0.122	1.701
Label propagation(n4)	0.531	0.034	1.553
Logistic network Lasso(n4)	0.536	0.034	**1.513**
Maximum likelihood(n8)	0.590	0.392	3.099
ML with edge detection(n8)	**0.431**	0.041	2.790
8-point median(n8)	0.544	0.182	3.737
Label propagation(n8)	0.564	0.055	1.883
Logistic network Lasso(n8)	0.561	0.055	1.811

With synthetic datasets, ML estimator with edge detection gave the best results. However, with 'edges+noise' dataset differences between different methods were small indicating that reliable edge detection was not possible due to noise. With 'edges+slope' dataset, maximum likelihood estimator with edge detection was the winner but semi-supervised learning based methods (label propagation and logistic network Lasso) followed closely. With real PV measurement data, the semi-supervised learning methods performed best. Presumably, the structure of discontinuities in 'edges+slope' dataset was oversimplified, and the semi-supervised methods performed better because they could utilize the global spatial structure of discontinuities in the real-life data.

VI. Related work

Several methods for outlier detection from spatial signals are reviewed in [1], [11]. Some methods model spatial data as multidimensional signal [2], other methods use graph signal processing [12], [13], [14]. However, indentification of edge-like discontinuities closely resembles image segmentation which has been studied a lot during the past decades [6]. In [2] a median-based algorithm was used for spatial outlier detection. [15] resembles our approach by solving outlier

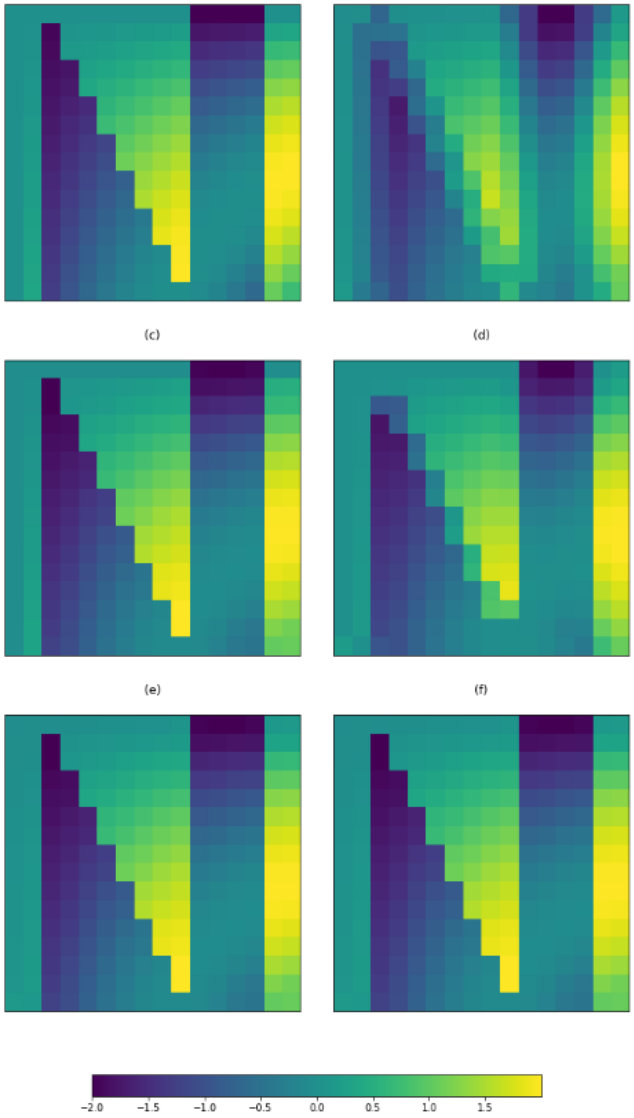

Fig. 8. Predictor results for the different algorithms: (a) original, (b) maximum likelihood, (c) ML with edge detection, (d) 4-point median (e) label propagation, (f) logistic network Lasso

detection problems with linear regression and regularization. They use vector-valued data points and also learn the weight vectors for logistic regression. Using vector dimension equal to one brings their approach close to our method based on logistic network Lasso. However, they identify the outliers directly while our approach uses logistic network Lasso for identifying discontinuities - which is just an intermediate step for constructing a predictor to be used in outlier detection.

Detection of partial shadows on PV systems have been studied in [16], [7], [17], but not for the detection of a single malfunctioning panel and not with the accuracy of predicting the power generated by individual panels.

VII. Conclusions

We compared different outlier detection methods for spatial data. Taking into account the edge-like discontinuities clearly improved the outlier detection performance. With PV panel monitoring data the best result were achieved using semi-supervised methods aiming at recognizing the global spatial pattern of discontinuities. In our previous research [18] utilizing convolutional neural networks for spatio-temporal outlier detection gave promising results. Our future research aims at comparing and combining these two approaches: explicitly estimating the spatial edge information and learning the irregular spatio-temporal nature from the data using artificial neural network based approaches.

Acknowledgment

We would like to thank SOLA Sense Ltd (www.solasense.fi) for providing the PV monitoring dataset for our research use.

References

[1] C. C. Aggarwal, *Outlier Analysis*, 2nd ed. Springer International Publishing, 2017.

[2] C. Lu, D. Chen, and Y. Kou, "Algorithms for spatial outlier detection," in *Third IEEE International Conference on Data Mining*, Nov. 2003, pp. 597–600.

[3] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, May 2013.

[4] I. Pitas and A. N. Venetsanopoulos, *Nonlinear digital filters: principles and applications*. Springer Science & Business Media, 2013, vol. 84.

[5] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. The MIT Press, 2010.

[6] R. M. Haralick and L. G. Shapiro, "Image Segmentation Techniques," in *Applications of Artificial Intelligence II*, vol. 0548, Apr. 1985, pp. 2–10.

[7] M. Bressan, Y. El-Basri, and C. Alonso, "A new method for fault detection and identification of shadows based on electrical signature of defects," in *2015 17th European Conference on Power Electronics and Applications (EPE'15 ECCE-Europe)*, Sep. 2015, pp. 1–8.

[8] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. Wiley, 2018.

[9] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Carnegie Mellon University, Technical Report CMU-CALD-02-107, 2002.

[10] H. Ambos, N. Tran, and A. Jung, "The Logistic Network Lasso," *arXiv:1805.02483 [cs, stat]*, May 2018.

[11] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 83:1–83:41, 2018.

[12] S. Shekhar, C.-T. Lu, and P. Zhang, "Detecting graph-based spatial outliers: algorithms and applications (a summary of results)," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 371–376.

[13] ——, "Detecting graph-based spatial outliers," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 451–468, 2002.

[14] ——, "A Unified Approach to Detecting Spatial Outliers," *GeoInformatica*, vol. 7, no. 2, pp. 139–166, Jun. 2003.

[15] M. Yamada, S. Liu, and S. Kaski, "Interpreting Outliers: Localized Logistic Regression for Density Ratio Estimation," *arXiv:1702.06354*, Feb. 2017.

[16] A. Woyte, J. Nijs, and R. Belmans, "Partial shadowing of photovoltaic arrays with different system configurations: literature review and field test results," *Solar Energy*, vol. 74, no. 3, pp. 217–233, Mar. 2003.

[17] F. Salem and M. A. Awadallah, "Detection and assessment of partial shading in photovoltaic arrays," *Journal of Electrical Systems and Information Technology*, vol. 3, no. 1, pp. 23–32, May 2016.

[18] T. Huuhtanen and A. Jung, "Predictive maintenance of photovoltaic panels via deep learning," in *2018 IEEE Data Science Workshop (DSW)*. IEEE, Jun. 2018, pp. 66–70.