Detecting Spatio-Temporal Outliers with Kernels and Statistical Testing

James P. Rogers
U.S. Army Engineer Research and Development Center
7701 Telegraph Road
Alexandria, VA 22315
Email: James.P.Rogers.II@erdc.usace.army.mil

Daniel Barbará
George Mason University
CS Department, MSN 4A5
Fairfax, VA 22030
Email: dbarbara@gmu.edu

Carlotta Domeniconi
George Mason University
CS Department, MSN 4A5
Fairfax, VA 22030
Email: carlotta@cs.gmu.edu

***Abstract*—Outlier detection is the discovery of points that are exceptional when compared with a set of observations that are considered normal. Such points are important since they often lead to the discovery of exceptional events. In spatio-temporal data, observations are vectors of feature values, tagged with a geographical location and a timestamp. A spatio-temporal outlier is an observation whose attribute values are significantly different from those of other spatially and temporally referenced objects in a spatio-temporal neighborhood. It represents an object that is significantly different from its neighbors, even though it may not be significantly different from the entire population. The discovery of outliers in spatio-temporal data is then complicated by the fact that one needs to focus the search on appropriate spatio-temporal neighborhoods of points. The work in this paper leverages an algorithm, StrOUD (Strangeness-based Outlier Detection algorithm), that has been developed and used by the authors to detect outliers in various scenarios (including vector spaces and non-vectorial data). StrOUD uses a measure of strangeness to categorize an observation, and compares the strangeness of a point with the distribution of strangeness of a set of baseline observations (which are assumed to be mostly from normal points). Using statistical testing, StrOUD determines if the point is an outlier or not. The technique described in this paper defines strangeness as the sum of distances to nearest neighbors, where the distance between two observations is computed as a weighted combination of the distance between their vectors of features, their geographical distance, and their temporal distance. Using this multi-modal distance measure (thereby called kernel), our technique is able to diagnose outliers with respect to spatio-temporal neighborhoods. We show how our approach is capable of determining outliers in real-life data, including crime data, and a set of observations collected by buoys in the Gulf of Mexico during the 2005 hurricane season. We show that the use of different weightings on the kernel distances allows the user to adapt the size of spatio-temporal neighborhoods.**

I. Introduction

Outlier detection is an important data mining task that deals with the discovery of points that are exceptional when compared with a set of observations that are considered "normal." These points are important since they often lead to the discovery of exceptional events. Outlier detection can reveal points that behave "anomalously" with respect to other observations. Examining such points can reveal clues to solve problems. In other cases, the sudden appearance of a large number of outliers can point to a change in the underlying process that is generating the data. Hawkins [6] defines an outlier as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism." This suggests the possibility of detecting outliers by knowing the "mechanism" by which the normal observations were generated and testing points for "membership" to this mechanism. Indeed, that is the path that early work in outlier detection followed (in the statistical community; see [9] for a comprehensive review): postulate a model for the probability distribution of normal points (e.g., a Gaussian model), and compute the likelihood of a point being generated by the postulated model. Unfortunately, coming up with the right model is, at best, as difficult as the original problem of finding outliers. This approach does not always work well in practice, because it can be seen as inducing a model over the normal data and using it to test points.

In spatio-temporal datasets, observations are tagged with their geographical location and their timestamp. These observations are a vector of attribute values for the features being measured. A spatio-temporal outlier is an observation whose values are significantly different from those of other spatially and temporally referenced objects in its spatio-temporal neighborhood. It represents an object that is significantly different from its neighborhood even though it may not be significantly different from the entire population.

The discovery of outliers in spatio-temporal data is complicated by needing to focus the search on an appropriate spatio-temporal neighborhood of points. The work in this paper leverages an algorithm, Strangeness-based Outlier Detection (StrOUD)[4], that has been developed and used to detect outliers in vector spaces and non-vectorial data. This paper will provide details on a technique capable of determining outliers in real datasets, such as a crime count dataset, a earthquake occurrence dataset, and a dataset of observations collected by buoys in the Gulf of Mexico during 2005.

A. Background

Recently, the field of statistical learning theory [10] has developed alternatives to *induction*: instead of using all the available points to induce a model, one can use the data (usually a small subset of it) to estimate unknown properties of points to be tested (e.g., membership to a class). This powerful idea leads to elegant algorithms that use standard statistical tests to compute the confidence on the estima-

tion. Using transduction, researchers have built Transductive Confidence Machines (TCM) (see [5]) which are able to estimate the unknown class of a point and attach confidence to the estimate. The transductive reliability estimation process has its theoretical foundations in the algorithmic theory of randomness developed by Kolmogorov [8]. Unlike traditional methods in machine learning, transduction can offer measures of reliability to individual examples, and uses very broad assumptions (it only assumes that the data points are independent and generated by the same stochastic mechanism). These properties make transduction an ideal mechanism to detect outliers, even though it has never been used before for that purpose. In [4], we have used the ideas of TCM to design a test that determines if a point is an outlier and attach a confidence to the estimation. New points are compared to a base of "normal" observations.

TCM [5] introduced the computation of the confidence using Algorithmic Randomness Theory [8]. (The first proposed application of Algorithmic Randomness Theory to machine learning problems, however, corresponds to Vovk et al. [11].) The confidence measure used in TCM is based upon universal tests for randomness, or their approximation. A Martin-Lof randomness deficiency test [8] based on such tests is a universal version of the standard p-value notion, commonly used in statistics. Martin-Lof proved that there exists a universal test for randomness smaller than any other test up to a multiplicative constant. Unfortunately, universal tests are not computable, and have to be approximated using non-universal tests called p-values. In the literature of significance testing, the p-value is defined as the probability of observing a point in the sample space that can be considered more extreme than a sample of data. This p-value serves as a measure of how well the data supports or does not support a null hypothesis. The smaller the p-value, the greater the evidence against the null hypothesis. Users of transduction as a test of confidence have approximated a universal test for randomness (which in its general form, is non-computable) by using a p-value function called *strangeness measure* [5] (or non-conformity score [11]). In truth, there is more than a single definition of strangeness measure, and in general, its definition depends on the base model used to construct the TCM. The general idea is that the strangeness measure corresponds to the uncertainty of the point being measured with respect to all the other labelled examples of a class: the higher the strangeness measure, means the higher the uncertainty.

In [4], we propose and use the following definition of strangeness (α), as follows: the strangeness α_i of a point i with respect to a baseline of data points is defined as shown in Equation 1, where the D_{ij} are the distances to the K closest neighbors of i in the baseline.

$$\alpha_i \;=\; \sum_{j=1}^{K} D_{ij} \tag{1}$$

This new definition will make the strangeness value of a point far away from points in the baseline considerably larger than the strangeness of points in the baseline. (This definition has been employed by [3] as a measure of isolation.) Using the α values, we can compute a p-value for a new point z_n, given a baseline of points $z_1, z_2, \cdots, z_{n-1}$ as given in Equation 2. Notice that in order to do that, the strangeness of each point z_j in the baseline has to be evaluated with respect to the baseline set $z_1, \cdots, z_{j-1}, z_{j+1}, \cdots, z_{n-1}$.

$$t(z_1, z_2, \cdots, z_n) \;=\; \frac{\#\{i = 1, \cdots, n:\ \alpha_i \;\geq\; \alpha_n\}}{n}. \tag{2}$$

The function $t()$ will measure the probability of having points in the baseline with strangeness greater than or equal to that of z_n. In general, a p-value is the maximum probability under the null hypothesis of the test statistic assuming a value equal to the observed outcome or a value just as extreme or more extreme (with respect to the direction indicated by alternative hypothesis) than the observed outcome. This gives us a way of testing the fitness of point z_n, by testing the null hypothesis H_0 as "z_n is fit to be in the baseline." Thus, the alternative hypothesis H_1 is "z_n is ill-fit to be in the basline." Selecting a confidence level $1 - \tau$ (usually 95 %), we can test if $p - value \;\leq\; \tau$, in which case, we reject all the null hypotheses and declare the point an outlier. Otherwise, we reject all the alternative hypotheses.

II. Spatio-Temporal StrOUD

To incorporate the effects of space and time, three kernels are embedded into the computation of distances for StrOUD. The total distance is the sum of the feature distance, spatial distance, and temporal distance, each multiplied by a positive factor that regulates the influence of each of the distances. The distance between two points is calculated using Equation 3. In the equation, d_f is the feature distance, d_s the spatial distance, and d_t the time distance, and $\rho + \beta + \delta = 1$

$$D(i,j) = \rho(d_f(i,j)) + \beta(d_s(i,j)) + \delta(d_t(i,j)) \tag{3}$$

In Equation 3, all the distance values $d()$ are normalized (using the distributions of feature, geographical, and time distances respectively), so they are constrained to the range $[0, 1]$. For a given i, the distances are only computed with respect to points j such that the timestamp of j precedes that of i (i.e., points that occur before i).

Given a dataset of observations, we apply the algorithm shown in Figure 1.

III. Experiments

A. Data

1) Crime Data: This data is a count of crimes that occurred in a geographic area. The data has one non-spatial attribute which is the count of the number of crimes. The data also has a spatial location (x coordinate and y coordinate) for each recorded count. This data set contains 464 instances. The location of the crimes is shown in Figure 2.

Given a dataset of m points
and a confidence level τ
For $i = 1$ to m do
    For every $j != i$ do
        Calculate and normalize $D(j,k)$ to every $k != i,j$ using Equation 3
        Compute α_j using Equation 1
        Compute $D(i,j)$ using Equation 3
    Compute α_i using Equation 1
    Compute the p_i as the p-value for i using Equation 2
    If $p_i < 1 - \tau$
    Declare i an outlier
End

Fig. 1. The Kernel-based StrOUD algorithm

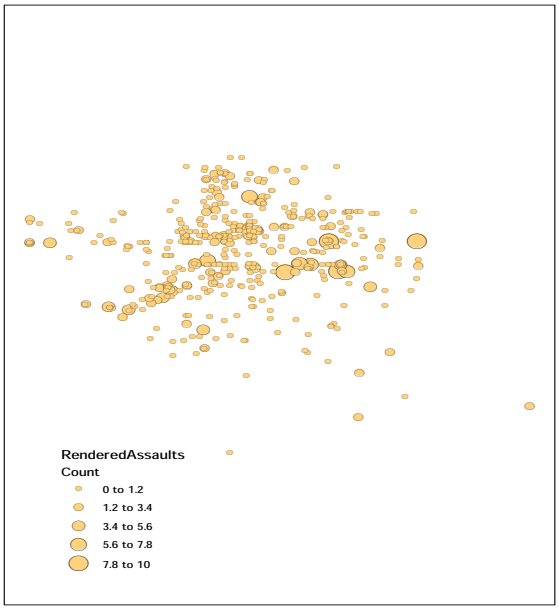

Fig. 2. Location of crimes

2) Earthquake Data from the southwestern United States: This data is for earthquakes that occurred in the southwestern United States. The data has two non-spatial attributes which are the earthquake's magnitude and depth. The data also has a spatial location (x coordinate and y coordinate) for each recorded earthquake. This data set contains 557 earthquake instances.

3) Buoy Data from the Gulf of Mexico: This data is weather data recorded from 30 buoys located in the Gulf of Mexico during 2005. The data has five non-spatial attributes which are wind direction, wind speed, barometric pressure, air temperature, and water temperature. Each buoy has a spatial location (x coordinate and y coordinate). This data set contains 204186 instances. Each instance lists the time (year, month, day, and hour) the attribute data was recorded at the spatial location. The location of the buoys in the Gulf of Mexico is shown in Figure 3.

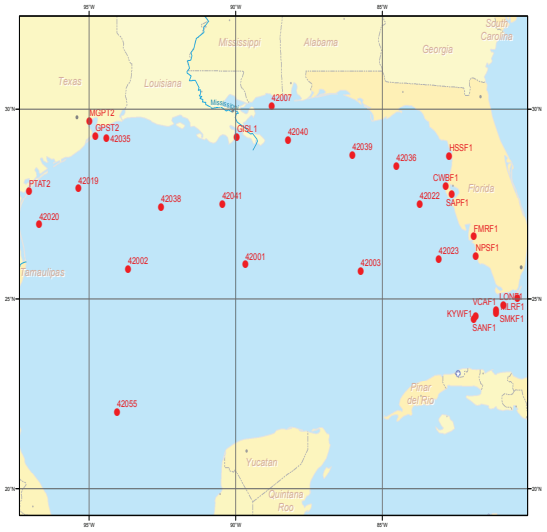

Fig. 3. Location of buoys in the Gulf of Mexico

B. Experimental Design

For the first experiments, the non-spatial attributes and the geographic information are utilized. Each individual geographic location and its non-spatial attributes for that location is an individual test point, and the remaining geographic locations with their non-spatial attributes are the training data. The data is looped through, so that each geographic point and its non-spatial attributes will be a test point once.

For the next experiments, the non-spatial attributes, the geographic information, and the temporal information are utilized. Each individual geographic location, the time the data was recorded, and its not-spatial attributes for that specific location and time is an individual test point. The remaining geographic locations with their non-spatial attributes and the time the data was recorded are the training data. The data is looped through, so that each point in the data set will be a test point once.

C. Crime Data Experiments

For the crime data, there are 464 instances. The minimum crime count value was 1 and the maximum value was 10. The weight for the non-spatial attributes decreases from 1 to 0 (1, .95, .75, .5, .25, .05, 0) while the weight of the spatial coordinates increases from 0 to 1(0, .05, .25, .5, .75, .95, 1).

The crime data experiment results were compared with the results of using Local Moran's module [1] in ArcGIS. The neighborhood in the Local Moran's analysis is the number of points within a radius of the point to be tested. 4000 feet was used as the radius for this experiment. The number of points within a 4000 foot radius of the test point were the number

of neighbors used in the Local Moran's calculations and the StrOUD calculations with the spatial kernel.

1) Crime Data Results: The algorithm finds 6 of the 464 instances were outliers when a weight of .95 was given to the count and a weight of .05 was given to the spatial coordinates. The location of the 6 outliers is shown in Figure 5. Local Moran's determined that 7 of the 464 instances were outliers. The location of the 7 outliers is shown in Figure 4. The point that Local Moran's determined to be an outlier and was not flagged by our method had a crime count of 3 and is located in a very dense crime count area. For this point, there are 174 of the 464 points within its 4000ft radius. Its z value (the deviation of the variable of interest with respect to the mean) is -2.56, and for a point to be an outlier with Local Moran's its z value must be less than -2.0, so this point is just below the threshold. In that neighborhood, 2.3% of the points have a count greater than 3, and 6.9% of the points have a count greater than or equal to 3.

D. Earthquake Experiments

For the earthquake data, there are 557 instances. For each instance, three experiments were conducted. The weight for the non-spatial attributes decreases from 1 to 0 (1, .98, .95, .75, .5, .25, .05, .02, 0) while the weight of the spatial coordinates increases from 0 to 1(0, .02, .05, .25, .5, .75, .95, .98, 1).

1) Earthquake Results: For each test case of the 557 instances, the script completed in approximately 10.5 seconds. The technique determined that data from 16 of the 557 sites were outliers. The results of the earthquake data experiments are shown in Figure 6. 12 of the 16 outliers were due primarily to the values of the non-spatial attributes and are shown in red in the figure. Two of the 16 outliers were due primarily to their spatial location and are shown in purple in the Figure. 1 of the 16 outliers was always an outlier regardless of the weights and is shown in orange in the figure. One of the 16 outliers needed contributions from both the non-spatial attributes and the spatial coordinates to be an outlier and is shown in blue in the Figure. This point was not determined to be an outlier when the weight of the non-spatial attributes or the spatial coordinates was equal to 1, but was determined to be an outlier when the weight of the non-spatial attributes varied between .95 and .25 and the weight of the spatial coordinates varied between .25 and .95. The remaining 541 sites were determined to not be outliers and are shown in green in the Figure.

E. Buoy Experiments

The first experiment was to run the buoy data with the original StrOUD algorithm using only the attributes (non-spatial and non-temporal). The second experiments ran the kernel-based StrOUD algorithm using only the attributes and spatial data ($\beta = 0$). The third experiments ran the kernel-based StrOUD script with the temporal kernel represented by a uniform distribution using the attributes and the temporal data. The fourth experiments ran the kernel-based StrOUD algorithm with all three components (attributes, spatial, and temporal).

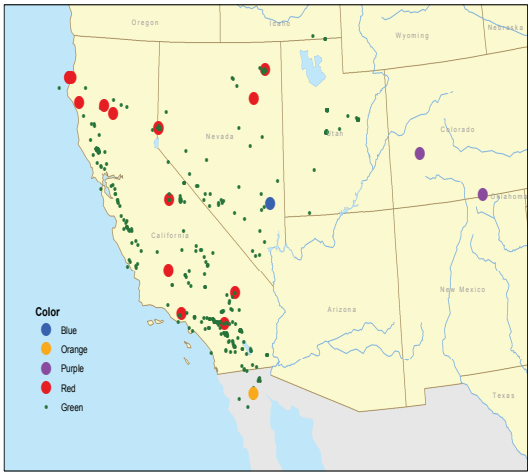

Fig. 6. Earthquake Data Results

1) Buoy Results: For the first experiment, a weight of 1 was given to the attributes and a weight of 0 was given to the spatial and temporal components, and 437 outliers were identified.

For the second set of experiments, the attributes and the spatial coordinates were weighted and the sum of their weights equaled 1, and a weight of 0 was given to the temporal component. For a weight of .95 for the attributes and a weight of .05 for the spatial coordinates, 373 outliers were identified, and all 373 outliers were in the group of 437 outliers found in the first experiment that did not include any spatial or temporal data. For a weight of .90 for the attributes and .10 for the spatial coordinates, 90 outliers were identified, and all 90 outliers were in the group of 373 outliers found when a weight .95 was given to the attributes and a weight of .05 was given to the spatial coordinates.

For the third set of experiments, a weight of 0 was given to the distance from spatial coordinates. For a weight of .95 for the attributes and a weight of .05 for the time, 401 outliers were identified, and 398 of those 401 outliers were in the group of 437 outliers found in the first experiment that did not include any spatial or temporal data. For a weight of .90 for the attributes and .10 for the temporal data, 110 outliers were identified, and all 110 outliers were in the group of 401 outliers found when a weight of .95 was given to the attributes and weight of .05 was given to time.

The fourth, and final, experiment was using the attribute information, the spatial coordinates, and the temporal information all together. For this experiment, a weight of .90 was given to the attributes, a weight of .05 was given to the spatial coordinates, and a weight of .05 was given to time. For this combination of weights, 123 outliers were identified. Of these 123 outliers, 120 of those outliers were in the group of 437 outliers found in the first experiment that did not include any

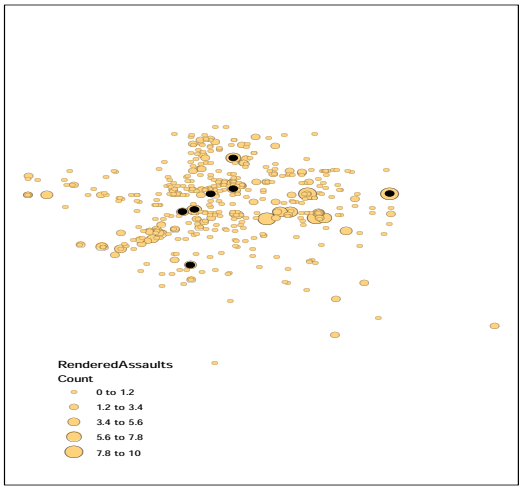

Fig. 4. Location of Outliers using Local Moran's

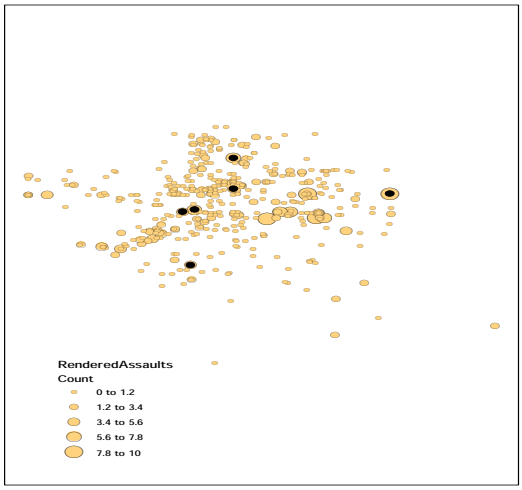

Fig. 5. Location of Outliers using StrOUD

spatial or temporal data, 120 of those outliers were in the group of 373 outliers found in the second experiment when a weight of .95 was given to the attributes and weight of .05 was given to the spatial coordinates, and all 123 outliers were in the group of 401 outliers found in the third experiment when a weight of .95 was given to the attributes and a weight of .05 was given to time.

Table I shows the results of the experiments with the buoy data. Except for three outliers found when a weight is given to the temporal kernel, all of the other outliers are in the set of 437 outliers resulting from the experiment when all the weight is given to the attributes and no weight is given to the spatial or temporal components. Also, all of the outliers resulting from a weight of .90 for the attributes and a total weight of .10 given to the spatial and temporal components are a subset of the outliers resulting from a weight of .95 for the attributes and a weight of .05 given to the spatial or temporal components. Strong outliers and weak outliers are defined such that strong outliers have a p-value between 0 and 0.02 and weak outliers have a p-value between 0.02 and 0.05. Giving weight to the spatial and temporal kernels in addition to the vector of features kernel resulted in identifying only the strong outliers. The identified outliers resulted from a hurricane or a strong cold front. Figure 7 shows the location of buoys producing weak outliers resulting from the location of Hurricane Katrina at 2am on August 29, 2005. Figure 8 shows the location of buoys producing both strong outliers and weak outliers resulting from the location of Hurricane Katrina at 7am on August 29, 2005. Figure9 shows the location of buoys producing outliers resulting from a cold front that moved across Texas into the Gulf of Mexico at 12pm on December 8, 2005.

TABLE I
RESULTS ON EXPERIMENTS WITH THE BUOY DATA

weights			Number of
attributes	spatial	time	Outliers
1.0	0	0	437
.95	.05	0	373
.90	.10	0	90
.95	0	.05	401
.90	0	.10	110
.90	.05	.05	123

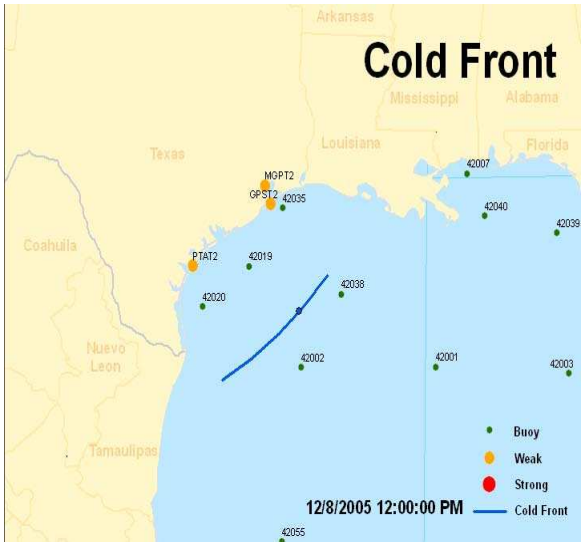

Fig. 9. Location of buoys with outliers resulting from a cold front

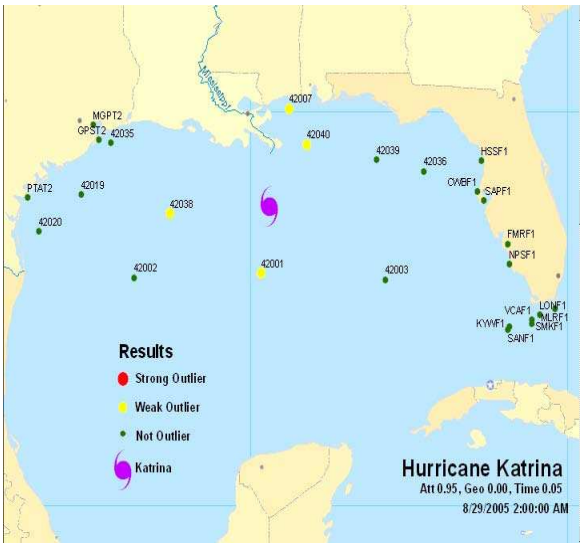

Fig. 7. Location of buoys with weak outliers resulting from Hurricane Katrina

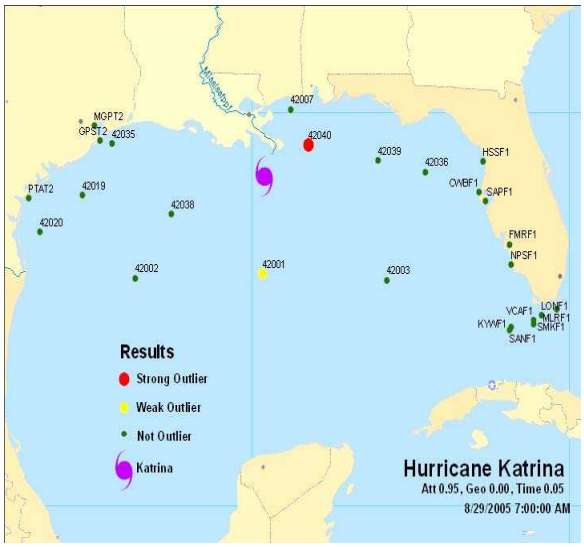

Fig. 8. Location of buoys with strong and weak outliers resulting from Hurricane Katrina

IV. Related Work

The work on [7] focuses on spatial outliers using a single attribute. Their methodology for finding outliers uses directed graph, neighborhoods, and comparing the attribute value to the average attribute value of its neighbors. This work, however, examines only one non-spatial attribute.

Local Moran's test [1], detects local spatial autocorrelation. It can be used to identify local clusters (regions where adjacent areas have similar values) or spatial outliers (areas distinct from their neighbors). However, the test is designed for univariate data (although, it has been extended to two dimensions).

In [2], the authors design an appropriate function that can effectively represent the outlierness of an object. The outlierness can be viewed as the difference between a single non-spatial attribute and its neighbors. Again, this work examines only one non-spatial attribute and requires the input of the number of requested outliers.

V. Conclusions

This paper presents a novel technique to detect outliers with data that has multiple attributes, spatial coordinates, and temporal information based on statistical testing, the application of transduction, kernels, and weighting. Spatial and temporal kernels were successfully added to StrOUD to account for the effects of space and time, in addition to the attributes (non-spatial and non-temporal), for statistically determining outliers.

We introduce a kernel for computing pairwise distances using the vector of attributes, spatial coordinates, and temporal component. These three components are weighted with factors whose sum equals to 1. Using this kernel, StrOUD diagnosed points as outliers with respect to their spatio-temporal neighborhoods.

Experiments were conducted on several datasets, with very good results. Changing the weights of the contributions of the non-spatial attributes and the weights of the contributions of the spatial coordinates has a significant impact on detecting outliers.

Acknowledgment

The authors would like to thank Guido Cervone for facilitating the buoys data.

References

[1] Anselin, L. (1995) Local indicators of spatial association - LISA. *Geographical Analysis*, **27(2)**, 93-115.

[2] Kou, Y., Lu, CT., and Chen, D. (2006) Spatial Weighted Outlier Detection. Proceedings of the SIAM Conference on Data Mining.

[3] Angiulli, F. and Pizzuti, C. (2005) Outlier mining in large high-dimensional data sets. IEEE Transactions on Knowledge and Data Engineering, 17(2): 203-215.

[4] Barbará, D., Domeniconi, C., and Rogers, J.P. (2006) Detecting Outliers using Transduction and Statistical Testing. *Proceedings of the twelve ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, August 20-23,*

[5] Gammerman, A., and Vovk, V. (2002) Prediction algorithms and confidence measures based on algorithmic randomness theory. Theoretical Computer Science. 287: 209-217.

[6] Hawkins, D. (1980) Identification of Outliers. Chapman and Hall, London.

[7] Shekhar, S., Lu, C.T., and Zhang, P. (2003) A Unified Approach to Spatial Outliers Detection, *Geoinformatica*, 7(2), June, 2003.

[8] Li, M., and Vitanyi, P. (1997) Introduction to Kolmogorov Complexity and its Applications. 2nd Edition, Springer Verlag.

[9] Lewis, B.V. (1994) Outliers in Statistical Data. John Wiley.

[10] Vapnik. V. (1998) Statistical Learning Theory, New York: Wiley.

[11] Vovk, V., Gammerman, A., and Saunders, C. (1999) Machine learning applications of algorithmic randomness. *Proceedings of the 16th Intl. Conference on Machine Learning.* 444-453.