Effective Outlier Detection using K-Nearest Neighbor Data Distributions: Unsupervised Exploratory Mining of Non-Stationarity in Data Streams

Kirk D. Borne (Department of Computational and Data Sciences, George Mason University, Fairfax, VA, kborne@gmu.edu, 703-993-8402 [voice], 703-993-9300 [fax])

Abstract:

We describe approaches and preliminary experiments that are aimed at monitoring and detecting change in self-monitored data streams. We introduce a new algorithm for outlier detection using K-Nearest Neighbor Data Distributions. We run experiments on a variety of data stream topologies and thereby demonstrate the effectiveness of the new algorithm in detecting outliers and in quantitatively estimating the outlyingness likelihood of anomalous data points in diverse data streams. These streams may occur in deep space probes or in autonomous robotic exploration probes in Earth environments. We first consider an illustrative example involving a spatially varying data stream, which we then invoke as an analogy to studies of temporal data streams. Any large dataset can benefit from a general approach in which parametric characterization and exploration of patterns in the data can be conducted with any independent variable (in time or parameter space). Temporal variations that are novel, unexpected, previously unknown, or outside the bounds of our existing classification schemes are scientifically worthy of further investigation, which may include an autonomous response from a remote sensing agent. We describe eigenstate monitoring in the context of high-throughput change-detection. Changes in the eigenstates of a system are essentially changes in the stationarity of the source. We describe the detection and characterization of non-stationarity through a variety of unsupervised learning algorithms. These can be used in space missions for the detection and characterization of new events, state (phase) transitions in monitored systems, onset of change points in science environments, and pattern drift in critical correlations.

Keywords:

Data Mining; Outlier Detection; Data Streams; Unsupervised Learning; Space Science

1 Introduction: Scientific Discovery across Heterogeneous Data Collections

New modes of discovery are enabled by the growth of data and computational resources in the sciences. This cyberinfrastructure includes databases, virtual observatories (distributed data), high-performance computing (clusters and petascale machines), distributed computing (the Grid, the Cloud, peer-to-peer networks), intelligent search and discovery tools, and innovative visualization environments (Eastman et al. 2005). Data streams from experiments, sensors, and simulations are increasingly complex and growing in volume. This is true in most sciences, including time-domain astronomical sky surveys, climate simulations, Earth observing systems, remote sensing image collections, and sensor networks. At the same time, we see an emerging confluence of new technologies and approaches to science, most clearly visible in the growing synergism of the four modes of scientific discovery: Sensors-Computing-Modeling-Data. This has been driven by numerous developments, including the information explosion, the development of dynamic intelligent sensor networks, the acceleration in high performance computing (HPC) power, and advances in algorithms, models, and theories. Among these, the most extreme is the growth in new data.

The acquisition of data in all scientific disciplines is rapidly accelerating and causing a nearly insurmountable data avalanche (Bell et al. 2007). Computing power doubles every 18 months (Moore's Law), corresponding to a factor of 100 in ten years. The I/O bandwidth (into and out of memory and databases) increases by 10% each year – a factor 3 in ten years. By comparison, data volumes approximately double every year (a factor of 1,000 in ten years). Consequently, as growth in data volume accelerates, especially in the natural sciences (where funding certainly does not grow commensurate with data volumes), we will fall further and further behind in our ability to access, analyze, assimilate, and assemble knowledge from our data collections – unless we develop and apply increasingly more powerful algorithms, methodologies, and approaches (Borne 2009b).

In the space and Earth sciences in particular, rapid advances in three technology areas (science facilities, detectors, and computation) have continued unabated (Gray & Szalay 2004), all leading to more data (Becla et al. 2006). In the sciences, the scale of data-capturing capabilities grows at least as fast as the underlying microprocessor-based measurement system (Gray et al. 2005). For example, in astronomy, the fast growth in CCD detector size and sensitivity has seen the average dataset size of a typical large astronomy sky survey project grow from hundreds of gigabytes 10 years ago (e.g., the MACHO survey), to tens of terabytes today (e.g., 2MASS and Sloan Digital Sky Survey [Brunner et al. 2001; Gray & Szalay 2004]), up to a projected size of tens of petabytes 10 years from now (e.g., LSST, the Large Synoptic Survey Telescope [Becla et al. 2006; Bell et al. 2007]). In survey astronomy, LSST will produce one 56Kx56K (3-Gigapixel) image of the sky every 20 seconds, generating nearly 30 TB of data every day for 10 years. In solar physics, NASA announced in 2008 a science data center specifically for the Solar Dynamics Observatory, which obtains one 4Kx4K image every 10 seconds, generating one TB of data per day. NASA recognizes that previous approaches to scientific data management and analysis will simply not work. We see the data flood in all sciences (e.g., numerical simulations, high-energy physics, bioinformatics, drug discovery, medical research, geosciences, climate monitoring and modeling) and outside of science (e.g., banking, healthcare, homeland security, retail marketing). The application of data mining, knowledge discovery, and e-discovery tools to

these growing data repositories is essential to the success of our social, financial, health, government, and scientific enterprises. This is an especially challenging scientific problem as all modern science disciplines will become even more data-intensive in the coming decade (Szalay, Gray, & VandenBerg 2002). Increasingly sophisticated computational and data science approaches will be required to discover the wealth of new scientific knowledge hidden within these new massive scientific data collections (Gray et al. 2002; Szalay et al. 2002; Borne 2006; Graham et al. 2007; Kegelmeyer et al. 2008; Borne 2009a). As illustrated schematically in Figure 1, scientific data collections are numerous and heterogeneous, as are the types of information extracted from the data, and as are the types of machine learning algorithms applied to the data collections in order to achieve scientific knowledge discovery in databases (KDD).

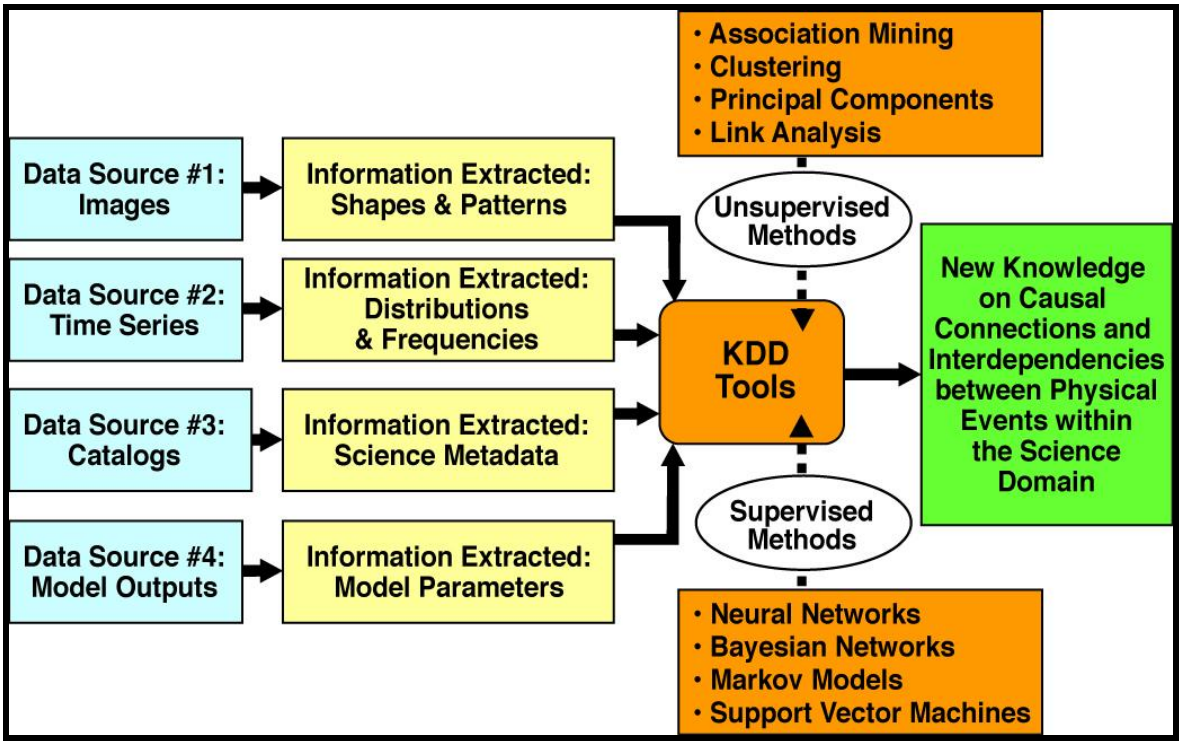

Figure 1 – The KDD process flow for scientific knowledge discovery.

Some of our prior results are reproduced in the following sections in order to demonstrate the application areas that we are investigating. In particular, Section 2 is reproduced, with slight modifications, from our SMC-IT 2006 paper (Borne 2006). In Section 3, Figures 3 and 4 and the paragraph describing those results are reproduced from our SDM 2009 paper (Das et al. 2009).

2 Unsupervised Exploration: Space Science Application

Unsupervised learning algorithms include principal components analysis (PCA), link analysis, self-organizing maps, association mining, and most clustering algorithms (including statistical clustering, such as mixture modeling). We begin our discussion with an example of an autonomous scientific data collection system operating as a science discovery machine in a remote environment with minimal or no human intervention. In this example, the operational behavior of the science data-collection system is data-driven, through machine learning (Borne

2001). Machine learning therefore provides decision support to the autonomous mission systems, in addition to providing decision support to human mission managers. The embedded machine learning algorithms are enhanced through data-sharing (from other sensors, spacecraft, databases, and/or models) – e-Science tools enable this data-sharing (Hey et al. 2002) – via Web Services, distributed data discovery and access, heterogeneous data fusion, distributed (Grid-like) model computations, and semantic data integration (Graham et al. 2007). We enumerate here a variety of machine learning applications for a hypothetical planetary rover. Numerous machine learning algorithms may be embedded within the roving operational *sciencecraft*:

(a) Supervised learning – search for rocks with known mineral compositions (by classifying each rock sample according to a known list of rock types).

(b) Unsupervised learning – discover objectively what types of rocks and minerals are present, without preconceived bias.

(c) Association mining – find the most common associations (co-occurrences) and also the most unusual co-occurrences of different minerals within rock samples.

(d) Clustering – find the complete set of unique classes of rocks.

(e) Classification – assign rock samples to known classes.

(f) Deviation/outlier detection – find one-of-a-kind, interesting, or anomalous rock/mineral samples.

(g) Learn as the rover goes from sample to sample – build up a model of the environment through Bayesian Networks or Markov modeling. Including spatial tools (such as GIS = Geographic Information Systems) to track the location of samples would provide still greater scientific insight and decision support capabilities.

(h) Information retrieval and fusion – relate the scientific instrument measurement results to other factors, such as dust storms, using data from other *sciencecraft* (e.g., from another rover, or from an orbiting satellite "mother ship").

(i) Decision trees and case-based reasoning – provide on-board intelligent data understanding and decision support (e.g., "stay here and do more" versus "move on to another rock;" or "send results to Earth immediately" versus "send results later").

(j) Case-based reasoning or logistic regression – predict where to go in order to find interesting rocks.

In all of these cases, decisions are based on the incoming data stream, prior experience, new knowledge, and decision logic. The rover can be allowed to act autonomously, without human intervention, in the deep space environment. Actions are determined by mining actionable data from all sensors. To maximize the decision-making accuracy and effectiveness, the rover should take advantage of other resources. These other resources may include measurements from other data-collecting sensors and models. The latter may be models of the environment (e.g., the geologic origins of the terrain, or the anticipated effects of an impending dust storm), or models

of the objects within the environment (e.g., location-dependent rock mineral classes), or models of the *sciencecraft's* behavior. These models can be updated in real-time as new data are acquired – this is data assimilation.

3 Unsupervised Exploration through Eigenstate Monitoring of Data Streams

We now consider the example of time-domain astronomy as an analogue to other space science data stream mining use cases (including machine learning in deep space missions, machine learning in sensor networks in Earth environments, and decision support in autonomous robotic exploration probes). We then argue how studies of large samples can benefit from a general approach in which parametric characterization and exploration of physical phenomena can be conducted with any independent variable (e.g., in time or space).

With time-domain astronomy, a new vision of the night sky will emerge, as time series are collected for billions of objects and thousands of object classes. Time variations may be detected in flux, in position, or in spectral properties. For those temporal variations that are novel, unexpected, previously unknown, or outside the bounds of our existing classification schemes, scientists will want to know (usually within seconds of the initial observation) that the event has occurred. Future projects may produce as many as 100,000 such event alerts each night for many years. The event alert notification must include as much information as possible to help the astronomers and facilities (especially robotic autonomous observatories) around the world to prioritize their response to each time-critical event. That information packet will include a probabilistic classification of the event, with some measure of the confidence of the classification (Bloom et al. 2008a, 2008b; Borne 2008a). Without good classification information in those alert packets, and hence without some means with which to prioritize the anticipated huge number of events from the new time-based instruments, the science community will consequently be buried in the data deluge and may miss some of the greatest scientific discoveries of the next decade.

To solve the anticipated massive event classification problem, the application of high-throughput change-detection algorithms is needed. These algorithms will use distributed astronomical databases worldwide to correlate in near real-time with each transient event, in order to model, classify, and prioritize correctly each event (Graham et al. 2007). In the seemingly simple case of variable stars, their variability is known, well studied, and well characterized already. However, if one of these stars' eigenmodes of variability were to change, then that would be extremely interesting – perhaps a signal that some potentially exotic astrophysical processes are taking place (Sarro et al. 2009). Scientists will definitely want to be notified promptly of these types of variations, which (in this case) are essentially *changes in the stationarity of the source*. Detection and characterization of non-stationarity can be measured through changes in the eigenvectors and eigenvalues of the light curve (Debosscher et al. 2007; Rebbapragada et al. 2009). Because these eigenstates (e.g., Fourier components) provide convenient and efficient short-hand representations of the features contained in the full data stream, the eigenstate information can be easily stored, monitored, and flagged as interesting and/or changing.

To address this type of parameter extraction and classification problem, new algorithms have been researched by our group (Giannella et al. 2006; Dutta et al. 2007; Das et al. 2009). These

algorithms are useful for distributed mining, change detection, and eigenvector monitoring in both static databases and dynamic data streams. Specifically, we investigated an eigenvector monitoring problem that addresses the research challenges associated with unsupervised exploratory mining of space science data streams. We analyzed the principal components of galaxy parameters as a function of an independent variable, similar to the temporal dynamic stream mining described above (which has time of observation as the independent variable). For our experiments, the independent variable was not the time of observation, but instead it was the local galaxy density, and it could have been any other major galaxy parameter (e.g., luminosity, color, concentration, effective radius, central surface brightness, metallicity).

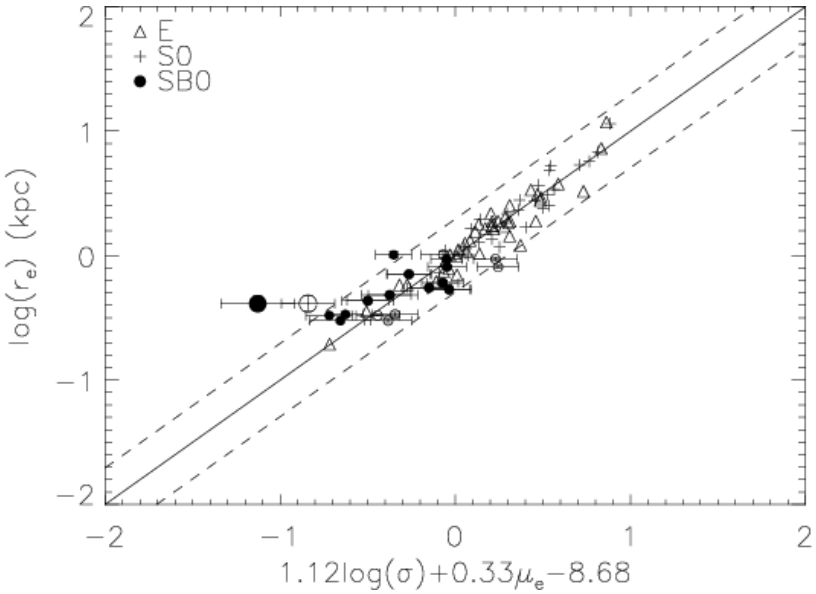

Figure 2 – Edge-on view of the fundamental plane (*continuous line*) of the Coma elliptical galaxies (*triangles*), S0 bulges (*open triangles*), and SB0 bulges (*filled circles).* The dashed lines represent the 3-σ deviation from the fundamental plane (reproduced from Aguerri et al. 2005).

The class of elliptical galaxies has been known for more than 20 years to show dimension reduction among a subset of physical attributes (radius, surface brightness, and central velocity dispersion), such that the 3-dimensional distribution of three of those astrophysical parameters reduces to a 2-dimensional plane (Djorgovski & Davis 1987). The normal to that plane represents the principal eigenvector of the distribution, and it is found that the first two principal components capture significantly more than 90% of the variance among those 3 parameters (e.g., see Figure 2; Aguerri et al. 2005).

By analyzing existing large astronomy databases, we have generated a sample of 102,600 galaxies. Each galaxy in this large sample was then assigned (labeled with) a new "local galaxy density" attribute, calculated through a volumetric Voronoi tessellation of the total galaxy distribution in space. The inverse of the Voronoi volume represents the local galaxy density (i.e., each galaxy occupies singly a well defined volume that is calculated by measuring the distance to its nearest neighbors in all directions and then generating the 3-dimensional convex polygon whose faces are the bisecting planes along the direction vectors pointing toward the nearest neighbors in each direction – the enclosed volume of the polygon is the Voronoi volume). It is

scientifically interesting to note that the dynamical timescale (age) of a gravitating system is proportional to the square root of the Voronoi volume (i.e., inversely proportional to the square root of the local galaxy density). Therefore, studying the variation of galaxy parameters and relationships as a function of the Voronoi volume is akin to studying the time evolution of the ensemble population of galaxies. In this way, the problem that we studied is dynamic in time.

For our initial correlation mining work, the entire galaxy data set was divided into 30 equal-sized partitions as a function of our independent variable: the local galaxy density. Consequently, each bin contains more than 3000 galaxies, thereby generating statistically robust estimators of the fundamental parameters in each bin. From the various astronomical data catalogs, we have extracted about four dozen measured parameters for each galaxy (out of a possible 800+ from the combined two catalogs): fluxes, size and radius measures, concentration indices, velocity dispersions, isophotal shape parameters, and surface brightness measures. For those parameters that depend on distance (e.g., radius), we have used the galaxy's measured redshift to normalize these into distance-independent physical measures for those parameters.

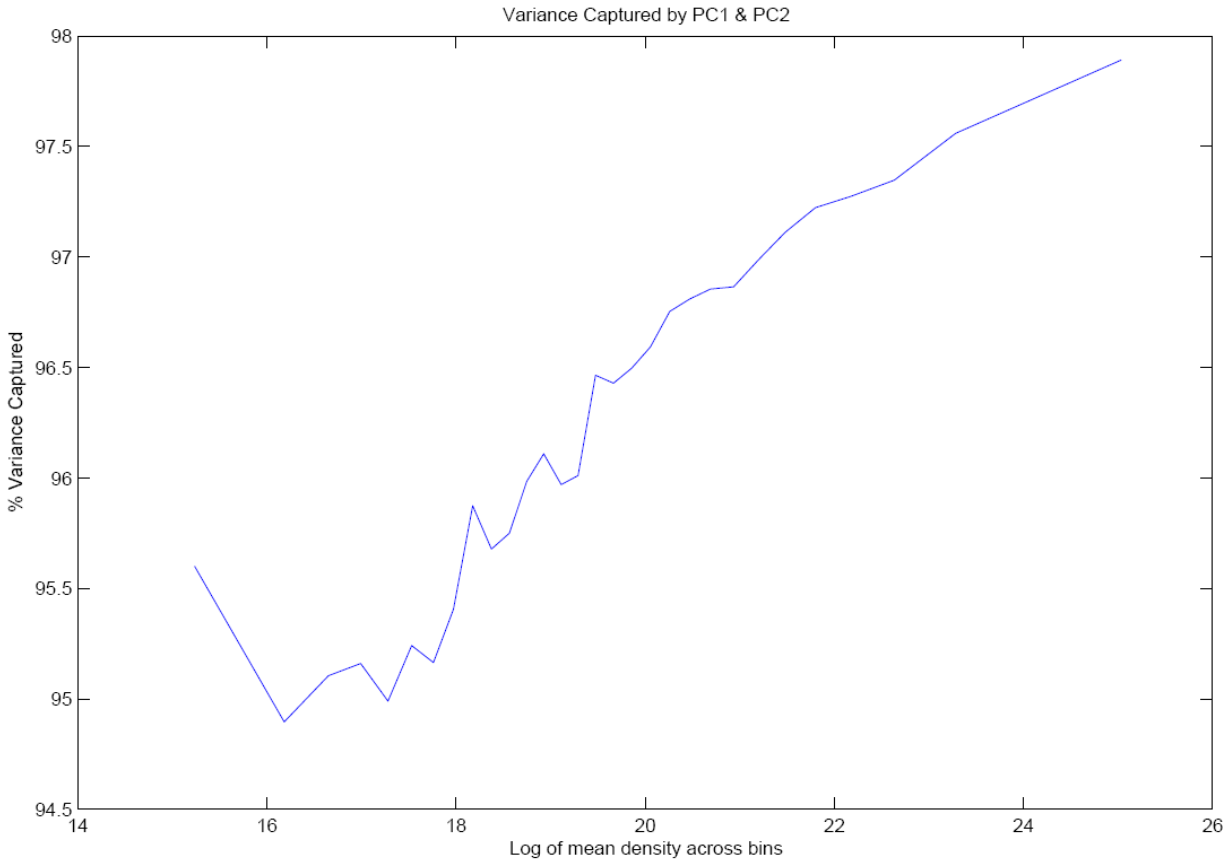

Figure 3 – Variance captured by the first two principal components of the fundamental plane as a function of the log of the mean galaxy local density (for 30 different bins containing ~3400 galaxies each). The sample parameters used in this analysis are i-band Petrosian radius containing 50% of the galaxy flux (from SDSS), velocity dispersion (SDSS), and K-band mean surface brightness (2MASS). This plot clearly shows that the fundamental plane relation becomes tighter with increasing local galaxy density (inverse Voronoi volume). (Borne et al. 2009)

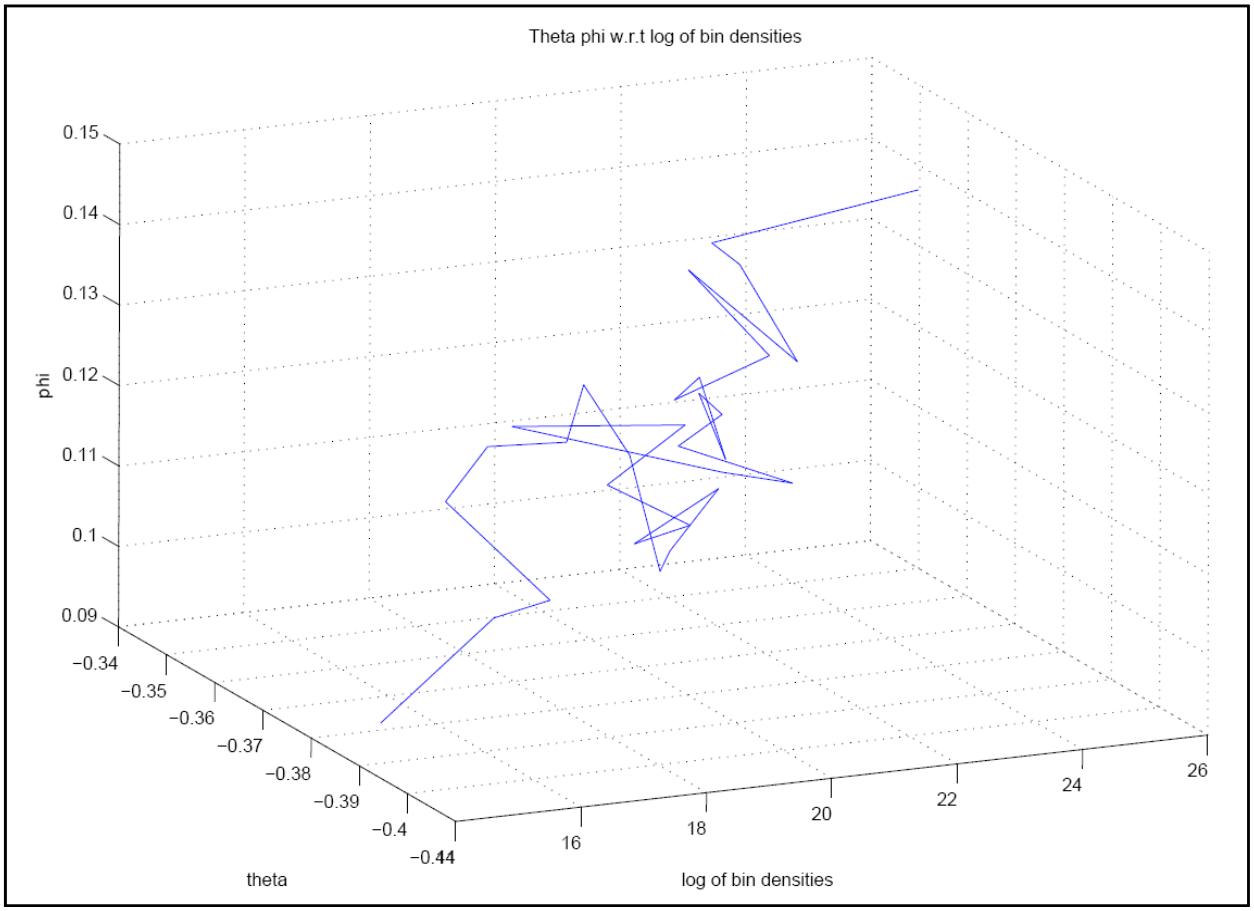

Figure 4 – Direction cosines for the normal vector to each fundamental plane point in Fig. 3, calculated as a function of local galaxy density (inverse Voronoi volume). Though there are some chaotic effects, in general there is a trend for the tilt of the fundamental plane to drift nearly systematically for elliptical galaxy ensembles ranging from low- to high-density regions (Das et al. 2009).

As a result of our data sampling criteria, we have been able to study eigenvector changes of the fundamental plane of elliptical galaxies as a function of density. Computing these eigenvectors for a very large number of galaxies, one density bin at a time, thus mimics the dynamic data stream challenge problem (stationarity or eigenvector change) described earlier. In addition, this galaxy problem actually has uncovered some new astrophysical results: we find that the variance captured in the first 2 principal components increases systematically from low-density regions to high-density regions (Figure 3), and we find that the direction of the principal eigenvector also drifts systematically in the 3-dimensional parameter space from low-density regions to the highest-density regions (Figure 4). The progression from low-density to high-density corresponds to an age progression from dynamically unevolved to dynamically evolved systems.

Eigenstate-monitoring tools such as this can be used in unsupervised exploratory data mining (either in dynamic data streams or in parameter sweeps in high-dimensional databases) for the detection of non-stationarity (e.g., new events, state/phase transitions, onset of change points, pattern drift in correlations). In addition, this could be a major enabling technology for machine learning in space science depending on the degree to which a data stream or a high-dimensional

data record can be accurately represented by its eigenvectors (principal components). This would lead to significant data and dimension reduction, and could be used to generate useful and scientifically meaningful condensed representations, summarizations, and abstractions of the data for more rapid analysis and transmission (Borne 2008b).

4 Related Work: Interestingness Detection - Discovery of Surprise, Novelty, and Anomalies

Interestingness detection refers to the discovery of novelty, outliers, anomalies, and surprise within large data sets and data streams. Novelty and surprise are two of the more exciting aspects of science – finding something totally new and unexpected – though perhaps it may represent a serious flaw, glitch, or error in a system. Petascale databases potentially offer a multitude of such opportunities. But how do we find that surprising novel thing? These come under various names: interestingness, outliers, novelty, surprise, anomalies, or defects (depending on the application). In large databases and in high-rate data streams, rapid detection and characterization of events (i.e., changes, outliers, anomalies, novelties) are essential. Various information theoretic measures of interestingness and surprise can be used for the task. Among these are Shannon entropy, information gain (Freitas 1998), Weaver's Surprise Index (1988), and the J-Measure (Smyth & Goodman 1991). In general, such metrics estimate the relative information content between two sets of discrete-valued attributes.

As an example, we note that rule learning algorithms, specifically decision tree rule induction, make use of the information gain metric in order to determine which attribute contains the most "information". This is the one attribute among all attributes that by itself yields the best single-attribute classifier (= the top-level decision node). After testing this attribute's value, the remaining attributes are tested again for the next best information gain. The ranking of each attribute's information gain provides a measure of attribute *interestingness*. When faced with a database of many hundreds of attributes (e.g., in a complex highly dimensional scientific data stream), the scientist end-user is unlikely to know in advance which attributes are most interesting. Consequently, the user usually ends up selecting the small handful of attributes that are most familiar to her/him. These may not be the most beneficial for scientific discovery or for efficient database exploration. Consequently, to address this problem, it is useful to analyze various measures of interestingness, including information gain, covariance analysis, PCA, and independent components analysis. The result will be an objective quantifiable feature-selection algorithm that presents the most interesting attributes to end-users for *efficient* and *effective* explorations: *efficient* in the sense that the selection of the most interesting attributes for query/retrieval avoids lots of useless searches and queries; *effective* in the sense that novel discoveries (beyond known classes and expected relationships) are enabled.

We have used PCA to identify outliers (Dutta et al. 2007, 2009). In particular, we have been studying cases where the first two PC vectors capture and explain most (>90%) of the sample variance in a data sample (specifically, the fundamental plane of elliptical galaxies). Consequently, in such a case, the component of a data record's attribute feature vector that projects onto the third PC eigenvector will provide a measure of the distance z_3 of that data record from the fundamental plane that defines the structure of the overwhelming majority of the data points. Simply sorting the records by z_3, and then identifying those with the largest values, will result in an ordered set of outliers (Dutta 2007) from most interesting to least interesting.

In many cases, the first test for outliers can be a simple multivariate statistical test of the "normalcy" of the data: is the location and scatter of the data consistent with a normal distribution in multiple dimensions? There are many tests for univariate data. Tests on multivariate data include the Shapiro-Wilk test for normalcy and the Stahel-Donoho multivariate estimator for outlyingness (Shapiro & Wilk 1965; Maronna & Yohai 1995). The Stahel-Donoho outlyingness parameter is straightforward to calculate and assign for each object:

$$\text{Outlyingness } O(x) = | x - \mu(X^n) | / \sigma(X^n) \text{ , where } x \in X^n = \{x_1, \ldots x_n\}$$

This is simply the absolute deviation of a data point x from the centroid of the data set X^n, normalized to the scale of the data. These tests should not be construed as proofs of non-normalcy or outlyingness, but as evidence. For machine learning of petascale data or for machine learning in deep space environments, even simple tests are non-trivial in terms of computational cost, but it is essential to apply a range of algorithms in order to make progress in mining the data. Our future work will investigate some of these techniques, including: *"Measures of Surprise in Bayesian Analysis"* (Bayarri & Berger 1997), *"Quantifying Surprise in Data and Model Verification"* (Bayarri & Berger 1998), and *"Anomaly Detection Model Based on Bio-Inspired Algorithm and Independent Component Analysis"* (Srinoy & Kurutach 2006).

Interestingness metrics are particularly useful as eigenstate-monitoring tools (for time series data streams and for parameter sweeps in high-dimensional databases) – these are used to detect new events, state (phase) transitions, onset of change points, non-stationarity, and pattern drift in correlations (Levy-Leduc & Roueff 2009).

5 Algorithm: Effective Outlier Detection using K-Nearest Neighbor Data Distributions

We have implemented a new algorithm for outlier detection that has proven to be effective at detecting a variety of novel, interesting, and anomalous data behaviors. The new algorithm evaluates the local data distribution around a test data point and compares that distribution with the data distribution within the sample defined by its K nearest neighbors. The algorithm's success is based on the assumption that the distribution of distances between a true outlier and its nearest neighbors will be different from the distribution of distances among those neighbors by themselves. This assumption relies on the definition of an outlier as a point whose behavior (i.e., the point's location in parameter space) deviates in an unexpected way from the rest of the data distribution. Our algorithm quantifies this deviation, and uses that quantity as a measure of $O(x)$, the "outlyingness" of the data point, or its "surprise index", or its "interestingness".

Our algorithm is different from the Distribution of Distances algorithm for outlier detection presented by Saltenis (2004), in which the comparison is between the local data distribution around a test data point and a uniform data distribution. Our algorithm is also different from the k-Nearest Neighbor Graph algorithm for outlier detection of Hautamaki et al. (2004), in which data points define a directed graph and outliers are those connected graph components that have just one vertex. Furthermore, our algorithm appears similar but is actually different in important ways from the incremental LOF (Local Outlier Factor) algorithms of Breunig et al. (2000) and Pokrajac et al. (2007), in which the outlier estimate is density-based (determined as a function of the data point's local reachability density), whereas our outlier estimate is based on the full local data distribution. Finally, we report that the KORM (K-median OutlieR Miner) approach to

outlier detection in dynamic data streams by Dhaliwal et al. (2010) is most similar to our algorithm, except that their approach is cluster-based (based on K-medians) whereas our approach is statistics-based. We describe our algorithm below.

For our algorithm, we define a general distance distribution function *f(d,x)* as the distribution of distances d between point x and a sample of data points. We specifically define $f_K(d,O)$ as the distribution of distances between a potential outlier O and its K-nearest neighbors. We also specifically define $f_K(d,K)$ as the distribution of distances among the K-nearest neighbors. Our algorithm compares the two distance distribution functions $f_K(d,O)$ and $f_K(d,K)$.

Our algorithm takes advantage of the two-sample K-S (Kolmogorov-Smirnov) statistical test, which is a classical non-parametric test used to estimate the likelihood that two sample distributions are drawn from the same population (= the Null Hypothesis). There is no assumption of normalcy or any other functional form for the distance distribution functions – this is an important and essential criterion in order to avoid introducing any bias in the estimation of outlier probability. We initially attempted to apply the Mann-Whitney (Wilcoxon) U Test to compare the two distance distribution functions, but this test failed to detect true outliers effectively – the primary reason is that the U Test essentially measures the difference in the median of the two distributions, which demonstrates that a single parameter (including the Stahel-Donoho outlyingness parameter *O(x)*, defined in Section 4) is often a completely insufficient indicator of true outlylingness. The p-value derived from the K-S statistic (= the maximum difference between the two samples' cumulative density functions) measures the likelihood that the two samples satisfy the Null Hypothesis (Wall 1996). We define an *Outlier Index* as (1-p) = the probability that the Null Hypothesis is invalid (i.e., that the data distributions are not drawn from the same population). Consequently, the *Outlier Index* measures the likelihood that the test data point deviates from the behavior of the remainder of the data stream. Our algorithm has the advantage that it makes no assumption about the shape of the data distribution or about "normal" behavior.

Algorithm – Outlier Detection using K-Nearest Neighbor Data Distributions

Find the set S(K) of K nearest neighbors to the test data point O.
Calculate the K distances between O and the members of S(K). These distances define $f_K(d,O)$.
Calculate the K(K-1)/2 distances among the points within S(K). These distances define $f_K(d,K)$.
Compute the cumulative distribution functions $C_K(d,O)$ and $C_K(d,K)$, respectively, for $f_K(d,O)$ and $f_K(d,K)$.
Perform the K-S Test on $C_K(d,O)$ and $C_K(d,K)$. Estimate the p-value of the test.
Calculate the *Outlier Index* = 1-p.
If *Outlier Index* $\geq$ 0.95, then mark O as an "Outlier". The Null Hypothesis is rejected.
If 0.90 < *Outlier Index* < 0.95, then mark O as a "Potential Outlier".
If p $\geq$ 0.10, then the Null Hypothesis is accepted: the two distance distributions are drawn from the same population. Data point O is not marked as an outlier.

According to this algorithm, an outlier is defined as a data point whose distribution of distances between itself and its K-nearest neighbors is measurably different from the distribution of distances among the K-nearest neighbors alone (i.e., the two sets of distances are not drawn from

the same population). We tested the effectiveness of this algorithm on a variety of synthetic idealized data streams. Our results are illustrated in Section 6.

6 Experimental Results

To test the K-Nearest Neighbor Data Distribution algorithm for outlier detection and its effectiveness, we performed a sequence of experiments on idealized data series. We synthesized three type of data streams:

a) Linear data streams
b) V-shaped data reversals (i.e., the "normal" data trend suddenly changes direction)
c) Circular-shaped data distributions

Then, we inserted test data points at varying distances from the "normal" data stream: from "true normal" to "soft outlier" to "hard outlier". We finally applied our algorithm and measured the *Outlier Index* for the test data points, which estimates the likelihood that the test points are outliers. In each experiment, there were 25 points in the data stream, from which we identified the K=9 nearest neighbors. Therefore, the 36 distances between these 9 points were calculated and used as an estimate for $f_K(d,K)$. Similarly, the 9 distances between the test data point and K nearest neighbors were calculated and used as an estimate for $f_K(d,O)$. In each of the scatter plots shown below (Figures 5, 7, and 9), the outlier is identified as a filled brown square, the K nearest neighbors are identified as filled green circles, and the remaining (non-nearest neighbor) points in the data stream are identified as filled blue diamonds.

6.1 Linear Data Streams

Figure 5 shows three sets of linear data streams (with noise) plus a test data point. In these tests, the outlier was shifted progressively from the middle of the stream (the "true normal" TN position) out to increasing distances above the stream line ("soft outlier" SO to "hard outlier" HO): these are labeled as Experiments L-TN, L-SO, and L-HO, respectively, in Table 1 (see section 6.4). An example of the comparison between the two cumulative distribution functions $C_K(d,O)$ and $C_K(d,K)$ used by the K-S Test is shown in Figure 6 for Experiment L-TN – this illustrates that the cumulative distributions are very similar, as they should be, for a non-outlier data point embedded in the middle of the stream's data distribution.

6.2 V-shaped Data Streams

Figure 7 shows three sets of V-shaped data streams (with noise) plus a test data point. In these tests, the outlier was shifted progressively from the apex of the V-shaped stream (the "true normal" position) out to increasing distances from the vertex in locations between the two branches of the V ("soft outlier" to "hard outlier"): these are labeled as Experiments V-TN, V-SO, and V-HO, respectively, in Table 1 (see section 6.4). An example of the comparison between the two cumulative distribution functions $C_K(d,O)$ and $C_K(d,K)$ used by the K-S Test is shown in Figure 8 for Experiment V-SO – this illustrates that the cumulative distributions are markedly different, as they should be, for an outlier data point outside the data distribution pattern of the stream.

6.3 Circular-shaped Data Streams

Figure 9 shows three sets of circular-shaped data streams (with noise) plus a test data point. In these tests, the outlier remained at the center of the circle while the inner radius of the circularly distributed data stream was shifted progressively outward from the central point (inner radius=0, the "true normal" position) out to increasing values for the inner radius, thus progressively isolating the test data point from the remainder of the data stream ("soft outlier" to "hard outlier"): these are labeled as Experiments C-TN, C-SO, and C-HO, respectively, in Table 1 (see section 6.4). An example of the comparison between the two cumulative distribution functions $C_K(d,O)$ and $C_K(d,K)$ used by the K-S Test is shown in Figure 10 for Experiment C-HO – this illustrates that the cumulative distributions are significantly different, as they should be, for an outlier data point that is distinctly removed from the data distribution pattern of the stream.

6.4 Outlier Index Determinations

Table 1 presents our experimental results: the KS Test p-value, the Outlier Index, and the Outlier Flag for the nine experiments described above. It is clear from this table that the K-Nearest Neighbor Data Distribution algorithm for outlier detection is very effective at identifying outliers and at quantitatively estimating their likelihood of "outlyingness". These results provide confidence that our new algorithm can be used to detect a variety of anomalous deviations in topologically diverse data streams.

Experiment ID	*Short Description of Experiment*	*KS Test p-value*	*Outlier Index = 1-p = Outlyingness Likelihood*	*Outlier Flag (p<0.05?)*
L-TN (Fig. 5a)	Linear data stream, True Normal test	0.590	41.0%	False
L-SO (Fig. 5b)	Linear data stream, Soft Outlier test	0.096	90.4%	Potential Outlier
L-HO (Fig. 5c)	Linear data stream, Hard Outlier test	0.025	97.5%	TRUE
V-TN (Fig. 7a)	V-shaped stream, True Normal test	0.366	63.4%	False
V-SO (Fig. 7b)	V-shaped stream, Soft Outlier test	0.063	93.7%	Potential Outlier
V-HO (Fig. 7c)	V-shaped stream, Hard Outlier test	0.041	95.9%	TRUE
C-TN (Fig. 9a)	Circular stream, True Normal test	0.728	27.2%	False
C-SO (Fig. 9b)	Circular stream, Soft Outlier test	0.009	99.1%	TRUE
C-HO (Fig. 9c)	Circular stream, Hard Outlier test	0.005	99.5%	TRUE

Table 1 Results of experiments on the effectiveness of the K-Nearest Neighbor Data Distribution algorithm for outlier detection.

7 Future Work

We will expand the experiments to test a variety of other outlier and anomaly detection algorithms, to compare these algorithms against the algorithm presented here, and to measure the precision and recall of outliers and anomalies from a diverse set of test data sets and realistic scientific data streams for each of the various algorithms. Effective outlier detection is not necessarily efficient, and vice versa. So we will examine the trade-offs between efficiency and effectiveness of the different algorithms for different types of data deviations, including changes in the trend line of the data, stochastic data variations, catastrophic data changes, changes in the median, median, and mode of the data stream, changes in the sample distribution, and more. Each of these types of deviations may have relevance in different space mission systems or in different sensor applications. Therefore, it is useful to determine which algorithms are most effective, and ultimately most efficient, in detecting non-stationary behavior in an unsupervised mode within remotely sensed data streams.

8 Summary: Machine Learning In Space

The acquisition of scientific data in all disciplines is accelerating and causing a nearly insurmountable data avalanche. Assimilating these data into models and using these data and models to drive scientific measurement systems are major scientific challenges for today's large scientific research projects. The application of machine learning algorithms will enhance the scientific return and knowledge-building capabilities of future space missions. Machine learning will enable the science missions to address data-intensive problems that would not otherwise be manageable. This will permit large scientific projects to make use of larger data volumes in the discovery and modeling process than is currently possible. The scientific return on the investments to build, launch, and operate future complex space missions will thus be maximized.

The challenges of real-time data analysis and exploration are growing as missions become increasingly complex in their instrumentation and as the missions produce exponentially more data in their telemetry packets, engineering streams, and science data systems. One such research challenge area is in the application of dynamic data-driven decision support in data-intensive environments, which we have discussed in this paper. In space science missions, the volume of data to be processed, analyzed, and explored and the corresponding demands on computational power thus lead naturally to an investigation of the information technology efficiencies that streamline the corresponding compute-intensive and data-intensive operations. Among these technologies are techniques that generate condensed representations of the data stream and knowledge extraction from data streams (e.g., eigenstate monitoring, or compressive sensing). The extraction of knowledge from the information content of a massive data collection is a clear example of data reduction. This “reduced” knowledge can be communicated and shared among instruments and missions in a much more bandwidth-friendly manner, which is essential for future in-orbit or deep-space data streams, which may be so voluminous that the data flow cannot be handled. Thus, the abstraction of data-intensive operations (through unsupervised learning applications, data mining models, and knowledge ontologies) may be the natural solution to the data-volume problem. For example, flight telemetry data streams may be processed and mined for non-stationarity (glitches, anomalies, and/or trajectory deviations), and could thus provide the necessary feedback to an intelligent systems loop that corrects the trajectory or other satellite

systems appropriately in real-time without human intervention. Similarly, flight engineering data streams can be mined for instrumentation problems or other hardware anomalies and, if possible, yield feedback to an on-board autonomous correction loop. Finally, though science data streams are projected to grow exponentially in volume, not all of these data need to be broadcast back to the ground, if appropriate distillations or summarizations are sufficient. It may be desired that on-board autonomous machine learning, data mining, and analysis systems perform preliminary processing steps and thus provide feedback to the science planning system – e.g., continue this observation, or stop and look elsewhere, or stop and send results to another spacecraft (such as a member of a constellation spacecraft system). In all of these cases, the data processing, data mining, information retrieval, and knowledge discovery processes are dynamic data-driven decision processes. Applying machine learning solutions to these data-intensive processes could have profound positive benefits for future space exploration.

Acknowledgments

We thank NASA for partial support of this work through the Applied Information Systems Research (AISR) program.

References:

Aguerri, J. A. L., et al., A&A, 434, 109 (2005)

Bayarri, M. J., & J. O. Berger: "Measures of Surprise in Bayesian Analysis," downloaded from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.6365 (1997)

Bayarri, M. J., & J. O. Berger: "Quantifying Surprise in the Data and Model Verification," downloaded from http://citeseer.ist.psu.edu/old/401333.html (1998)

Becla, J., et al.: Designing a Multi-petabyte Database for LSST. arXiv:cs/0604112v1 (2006)

Bell, G., Gray, J., & Szalay, A.: Petascale Computational Systems. arXiv:cs/0701165v1 (2007)

Bloom, J. S., et al., Astronomische Nachrichten, 329, 284 (2008a)

Bloom, J. S., Butler, N. R., & Perley, D. A., in Gamma-Ray Bursts 2007, AIP Conference Proceedings, vol. 1000, p. 11 (2008b)

Borne, K.: Science User Scenarios for a VO Design Reference Mission: Science Requirements for Data Mining, in Virtual Observatories of the Future, p.333 (2001)

Borne, K. D.: Data-Driven Discovery through e-Science Technologies. 2nd IEEE Conference on Space Mission Challenges for Information Technology (SMC-IT) (2006)

Borne, K.: A Machine Learning Classification Broker for the LSST Transient Database, Astronomische Nachrichten, vol. 329, p. 255 (2008a)

Borne, K.: Data Science Challenges from Distributed Petascale Astronomical Sky Surveys, in the DOE Workshop on Mathematical Analysis of Petascale Data, downloaded from http://www.orau.gov/mathforpetascale/slides/Borne.pdf (2008b)

Borne, K.: Scientific Data Mining in Astronomy, in Next Generation Data Mining, Chapman & Hall, pp. 91-114 (2009a)

Borne, K.: Astroinformatics: A 21st Century Approach to Astronomy. arXiv:0909.3892v1 (2009b)

Borne, K. D., Vedachalam, A., & Giannella, C. 2009, in preparation.

Breunig, M., Kriegel, H.-P., Ng, R., & Sander, S.: "LOF: Identifying Density-Based Local Outliers," ACM SIGMOD Record, Vol. 29, pp. 93-104 (2000)

Brunner, R., et al.: Massive Datasets in Astronomy. arXiv:astro-ph/0106481v1 (2001)

Das, K., Bhaduri, K., Arora, S, Griffin, W., Borne, K., Giannella, C., & Kargupta, H., "Scalable Distributed Change Detection from Astronomy Data Streams using Eigen-monitoring Algorithms", in SIAM Data Mining (SDM) (2009)
Debosscher, J., et al., A&A, 475, 1159 (2007)
Dhaliwal, P., Bhatia, M., & Bansal, P.: "A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: K-median OutlieR Miner)", Journal of Computing, vol. 2, pp. 74-80 (2010)
Djorgovski, S. G., & Davis, M., ApJ, 313, 59 (1987)
Dutta, H.: "Empowering Scientific Discovery by Distributed Data Mining on the Grid Infrastructure," Ph.D. dissertation, UMBC (2007)
Dutta, H., C. Giannella, K. Borne, & H. Kargupta, H.: "Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System," in the peer-reviewed proceedings of the 2007 SIAM International Conference on Data Mining (2007)
Dutta, H., C. Giannella, K. Borne, H. Kargupta, & R. Wolff: "Distributed Data Mining for Astronomy Catalogs," submitted to the IEEE Transactions in Knowledge and Data Engineering (2009a)
Eastman, T., Borne, K., Green, J, Grayzeck, E., McGuire, R., & Sawyer, D.: eScience and Archiving for Space Science. Data Science Journal, 4, 67-76 (2005)
Freitas, A.: "On Objective Measures of Rule Surprisingness," LNCC, 1510, pp. 1-9 (1998)
Giannella, C., H. Dutta, K. Borne, R. Wolff, & H. Kargupta: "Distributed Data Mining for Astronomy Catalogs," in the peer-reviewed proceedings of the 9th Workshop on Mining Scientific and Engineering Datasets, SIAM International Conference on Data Mining (2006)
Graham, M., Fitzpatrick, M., & McGlynn, T.: The National Virtual Observatory: Tools and Techniques for Astronomical Research, ASP Conference Series, Vol. 382 (2007)
Gray, J., et al.: arxiv.org/abs/cs/0202014 (2002)
Gray, J., & Szalay, A.: Where the Rubber Meets the Sky: Bridging the Gap between Databases and Science. Microsoft technical report MSR-TR-2004-110 (2004)
Gray, J., et al.: Scientific Data Management in the Coming Decade, arxiv.org/abs/cs/0502008 (2005)
Hautamaki, V. , Karkkainen, I., & Franti, P.: "Outlier Detection Using k-Nearest Neighbour Graph," Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04) (2004)
Hey, J., & Trefethen, A.: The UK e-Science Core Programme and the Grid, Future Generation Computer Systems, 18, 1017-1031 (2002)
Kegelmeyer, P., et al.: Mathematics for Analysis of Petascale Data: Report on a DOE Workshop, http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/PetascaleDataWorkshopReport.pdf (2008)
Levy-Leduc, C., & F. Roueff: "Detection and Localization of Change-Points in High-Dimensional Network Traffic Data," Annals of Applied Statistics, 3, 637 (2009)
Maronna, R. A. & V. J. Yohai: "The Behavior of the Stahel-Donoho Robust Multivariate Estimator," Journal of the American Statistical Association, 90, 330 (1995)
Pokrajac, D., Lazarevic, A., & Latecki, L.: "Incremental Local Outlier Detection for Data Streams," in IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (2007)
Rebbapragada ,U., et al.: "Finding Anomalous Periodic Time Series: An Application to Catalogs of Periodic Variable Stars," arXiv.org:0905.3428v1 (2009)
Saltenis, V.: "Outlier Detection Based on the Distribution of Distances between Data Points," Informatica, Vol. 15, No. 3, pp. 399-410 (2004).
Sarro, L., et al., A&A, 494, 739 (2009)
Shapiro, S. S., & M. B. Wilk: "An analysis of variance test for normality (complete samples)," Biometrika, 52, pp. 591–611 (1965)
Smyth, P., & R. M. Goodman: "Rule Induction Using Information Theory," in Knowledge Discovery in Databases, pp 159-176, AAAI/MIT Press (1991)
Srinoy, S., & W. Kurutach: "Anomaly Detection Model Based on Bio-Inspired Algorithm and Independent Component Analysis," TENCON 2006, IEEE Region 10 Conference proceedings, pp. 1-4 (2006)

Szalay, A., et al.: Petabyte Scale Data Mining: Dream or Reality? in Proceedings of the SPIE, volume 4836, Survey and Other Telescope Technologies and Discoveries, p. 333 (2002)
Szalay, A. S., Gray, J., & VandenBerg, J.: Petabyte Scale Data Mining: Dream or Reality? arXiv:cs/0208013v1 (2002)
Wall, J. V.: "Practical Statistics for Astronomers: Correlation, Data-Modeling, and Sample Comparison," Quarterly Journal of the Royal Astronomical Society, Vol. 37, pp. 519-563 (1996)
Weaver's Surprise Index: Encyclopedia of Statistical Sciences, Vol. 9, pp. 104-109, Wiley (1988)

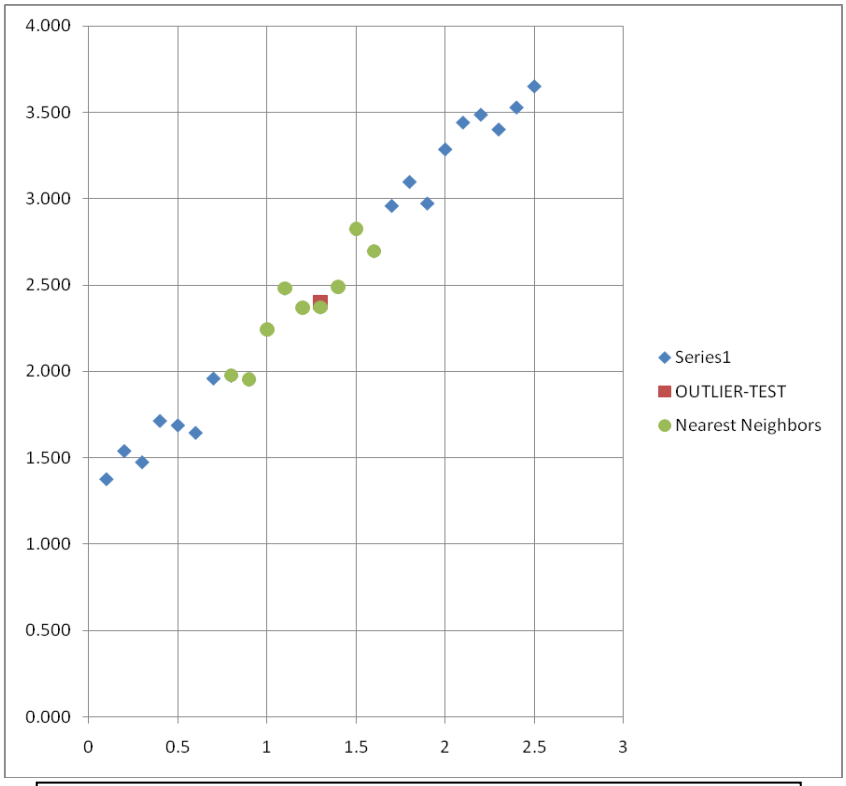

Figure 5a – Experiment L-TN (see Section 6.1 and Table 1)

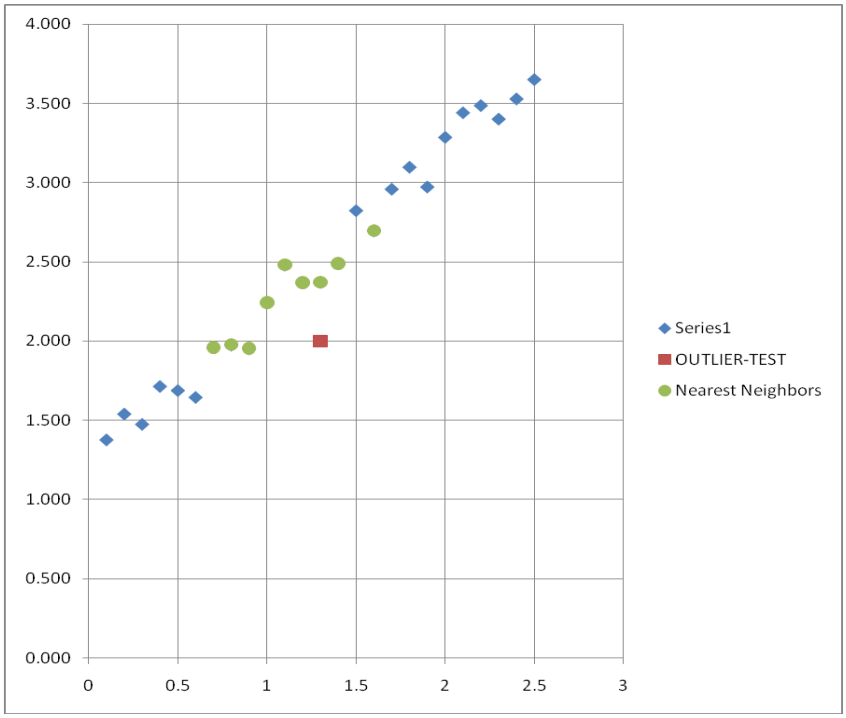

Figure 5b – Experiment L-SO (see Section 6. 1 and Table 1)

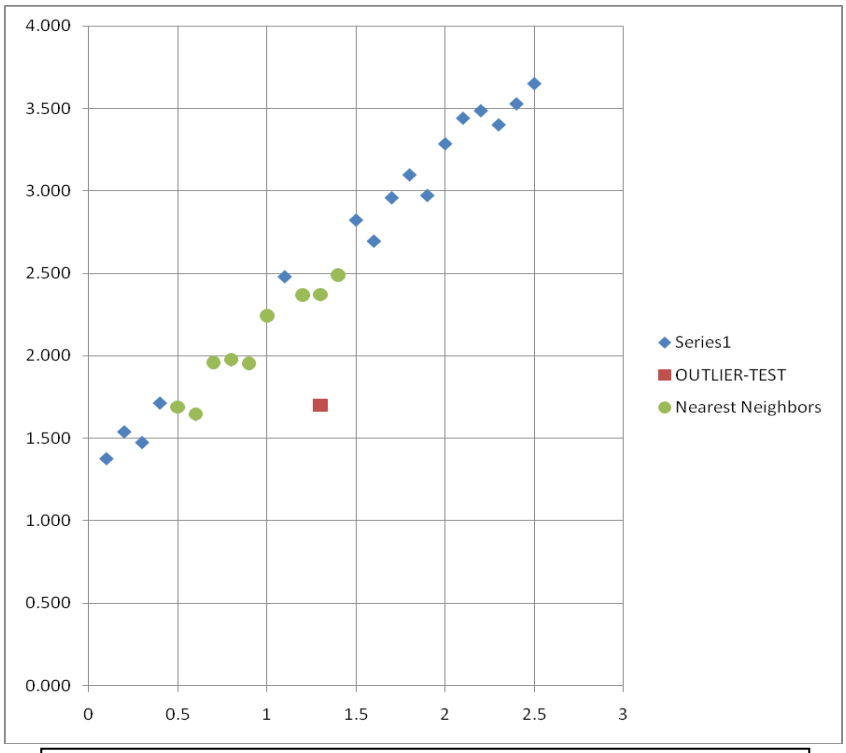

Figure 5c – Experiment L-HO (see Section 6. 1 and Table 1)

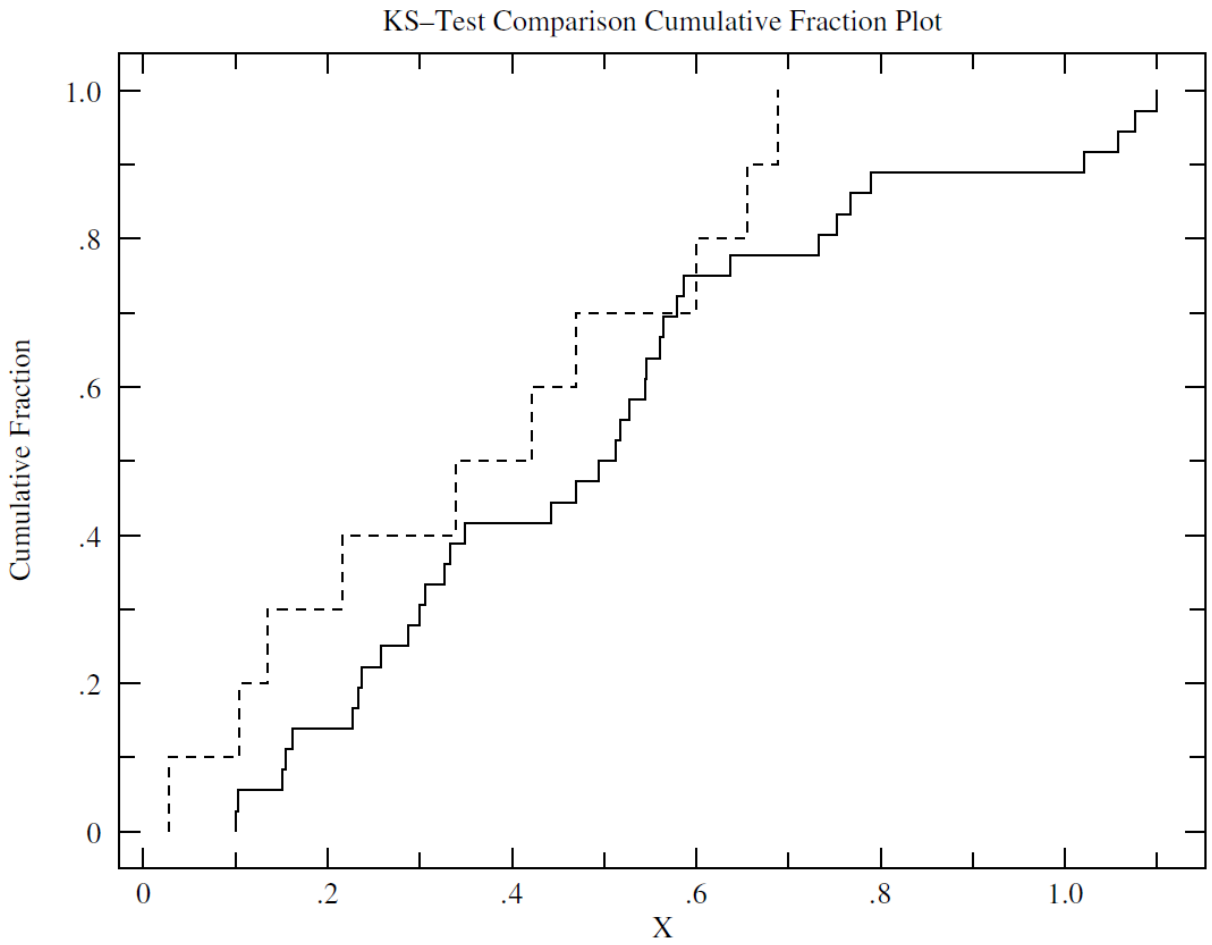

Figure 6 – Cumulative distribution plots for Experiment L-TN (see Section 6.1) used in the KS-Test to test if the two distance distribution functions $f_K(d,O)$ and $f_K(d,K)$ are drawn from the same population (Section 5). The ordinate X refers to the distance between data points in the sample. The solid line is the cumulative distribution of distances exclusively among the K nearest neighbors to the test data point. The dashed line is the cumulative distribution of distances between the data point and its K nearest neighbors (illustrated in Figure 5a).

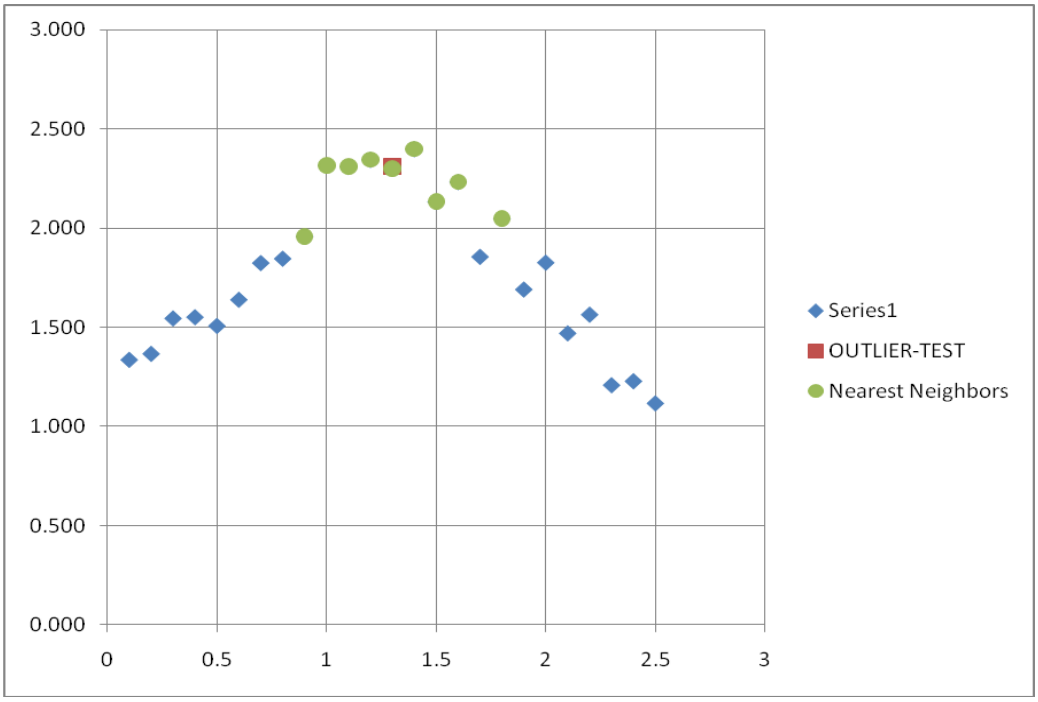

Figure 7a – Experiment V-TN (see Section 6.2 and Table 1)

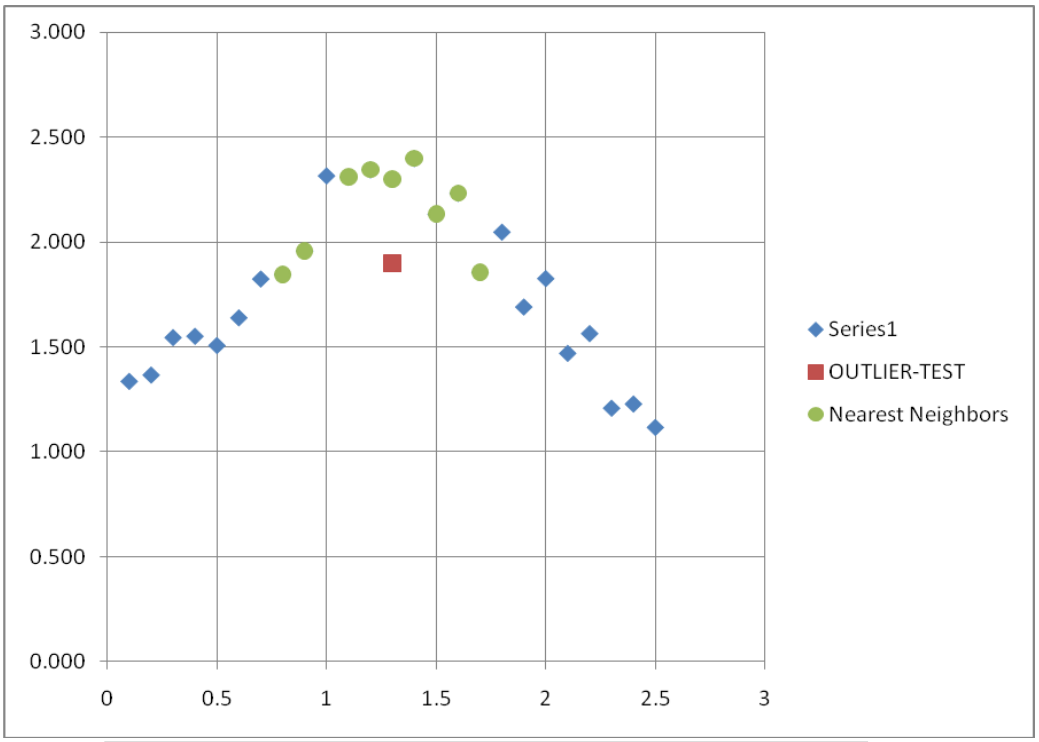

Figure 7b – Experiment V-SO (see Section 6.2 and Table 1)

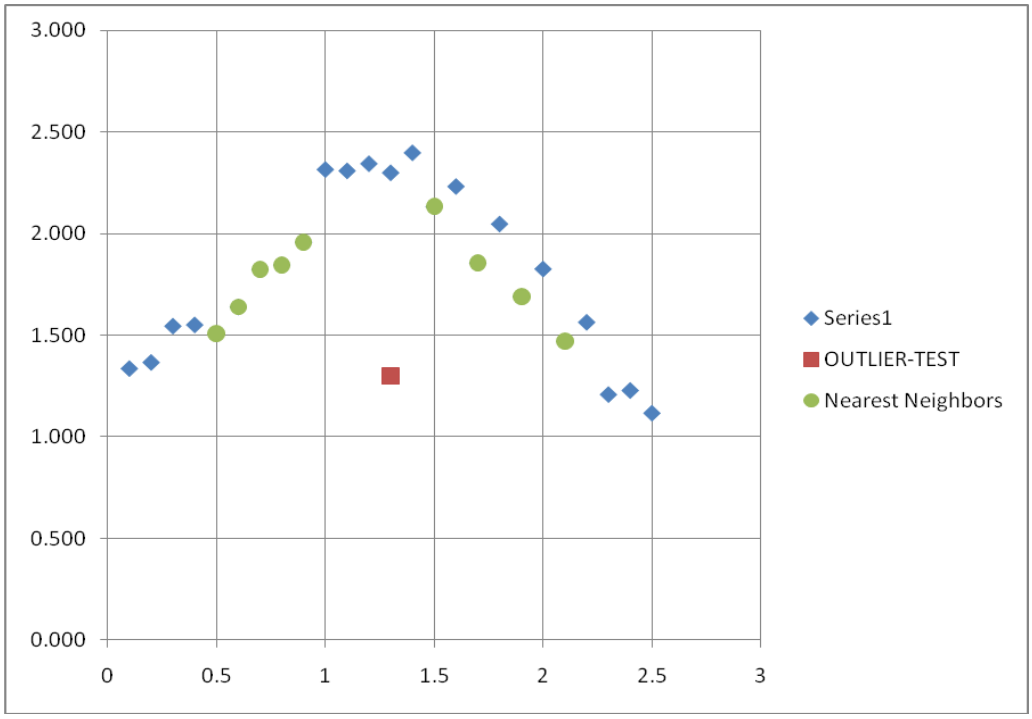

Figure 7c – Experiment V-HO (see Section 6.2 and Table 1)

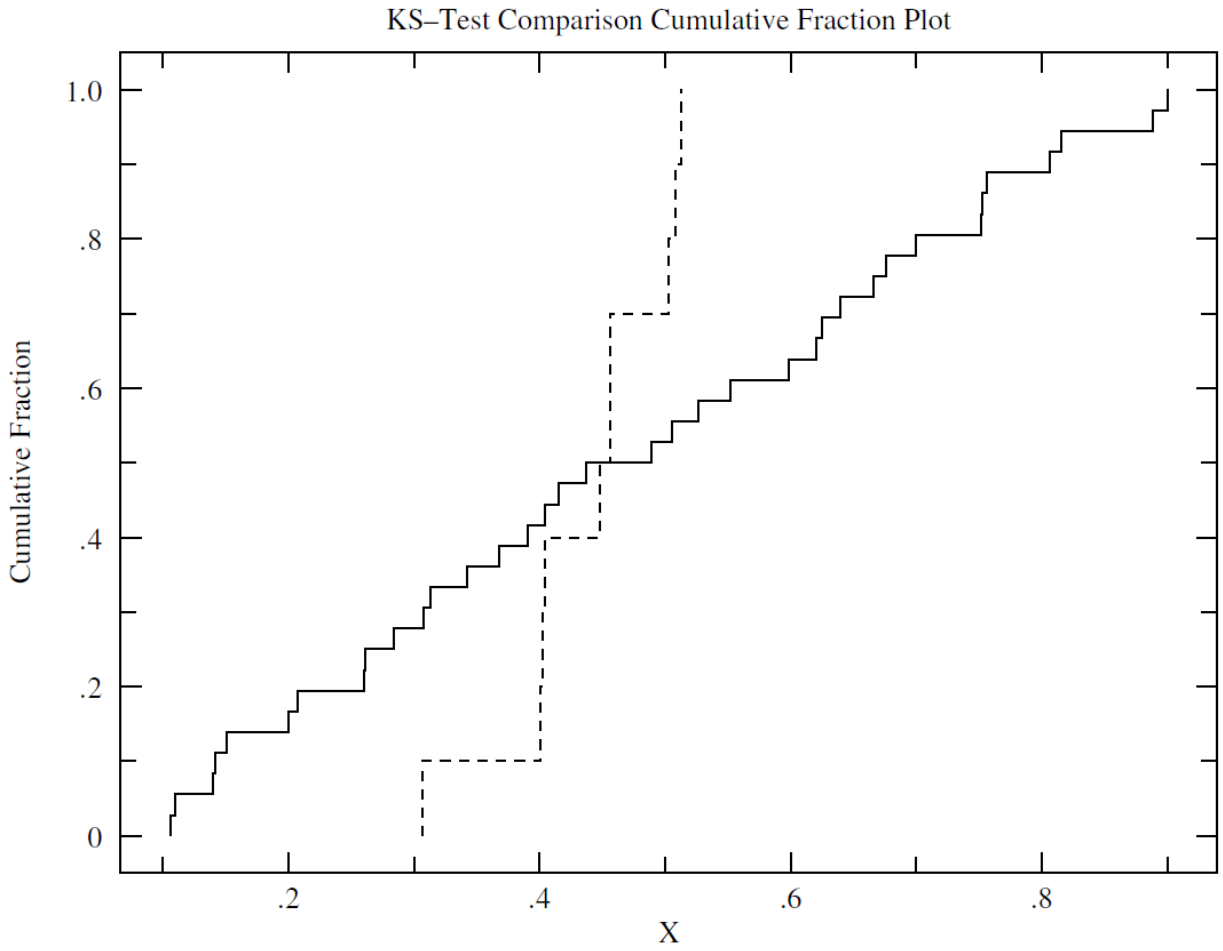

Figure 8 – Cumulative distribution plots for Experiment V-SO (see Section 6.2) used in the KS-Test to test if the two distance distribution functions $f_K(d,O)$ and $f_K(d,K)$ are drawn from the same population (Section 5). The ordinate X refers to the distance between data points in the sample. The solid line is the cumulative distribution of distances exclusively among the K nearest neighbors to the test data point. The dashed line is the cumulative distribution of distances between the data point and its K nearest neighbors (illustrated in Figure 7b).

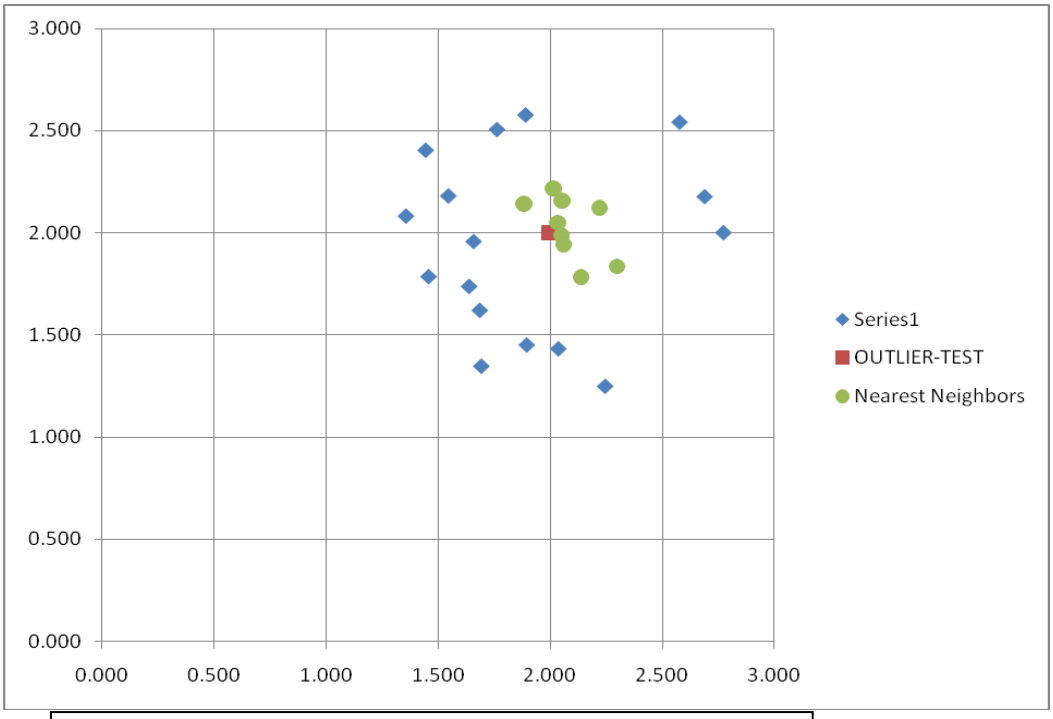

Figure 9a – Experiment C-TN (see Section 6.3 and Table 1)

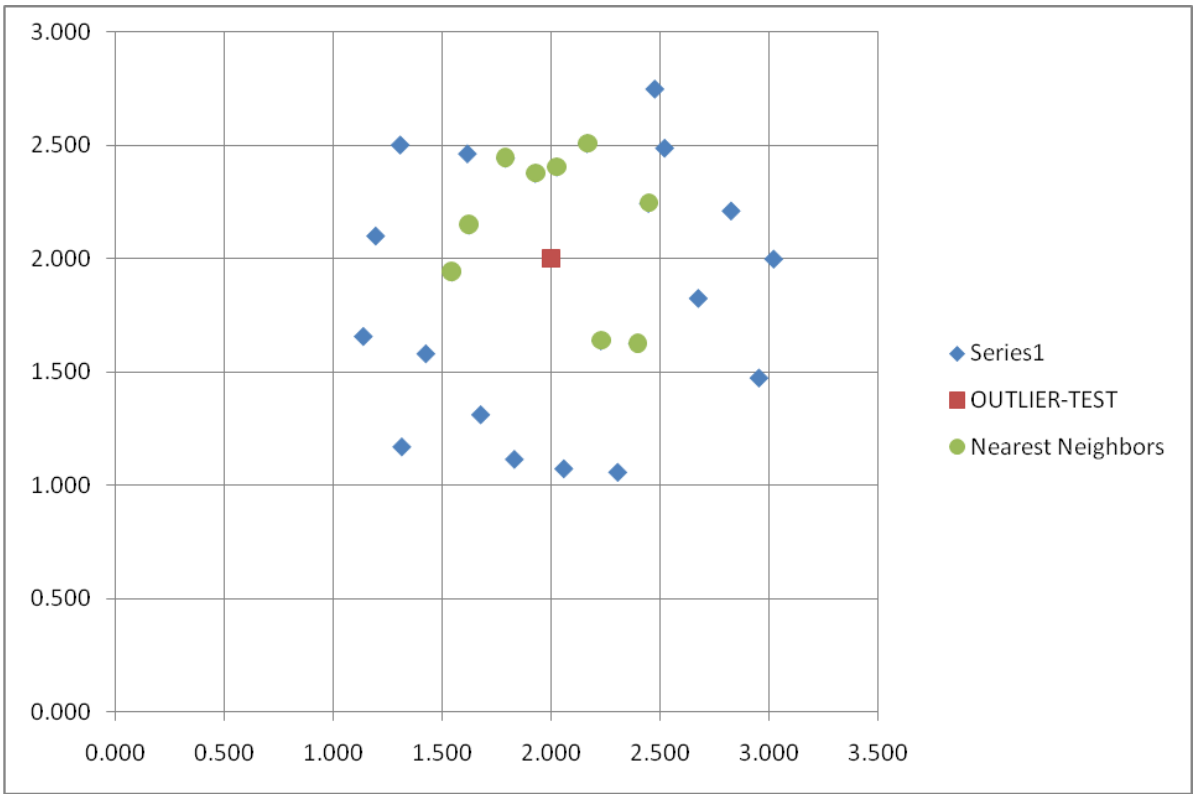

Figure 9b – Experiment C-SO (see Section 6.3 and Table 1)

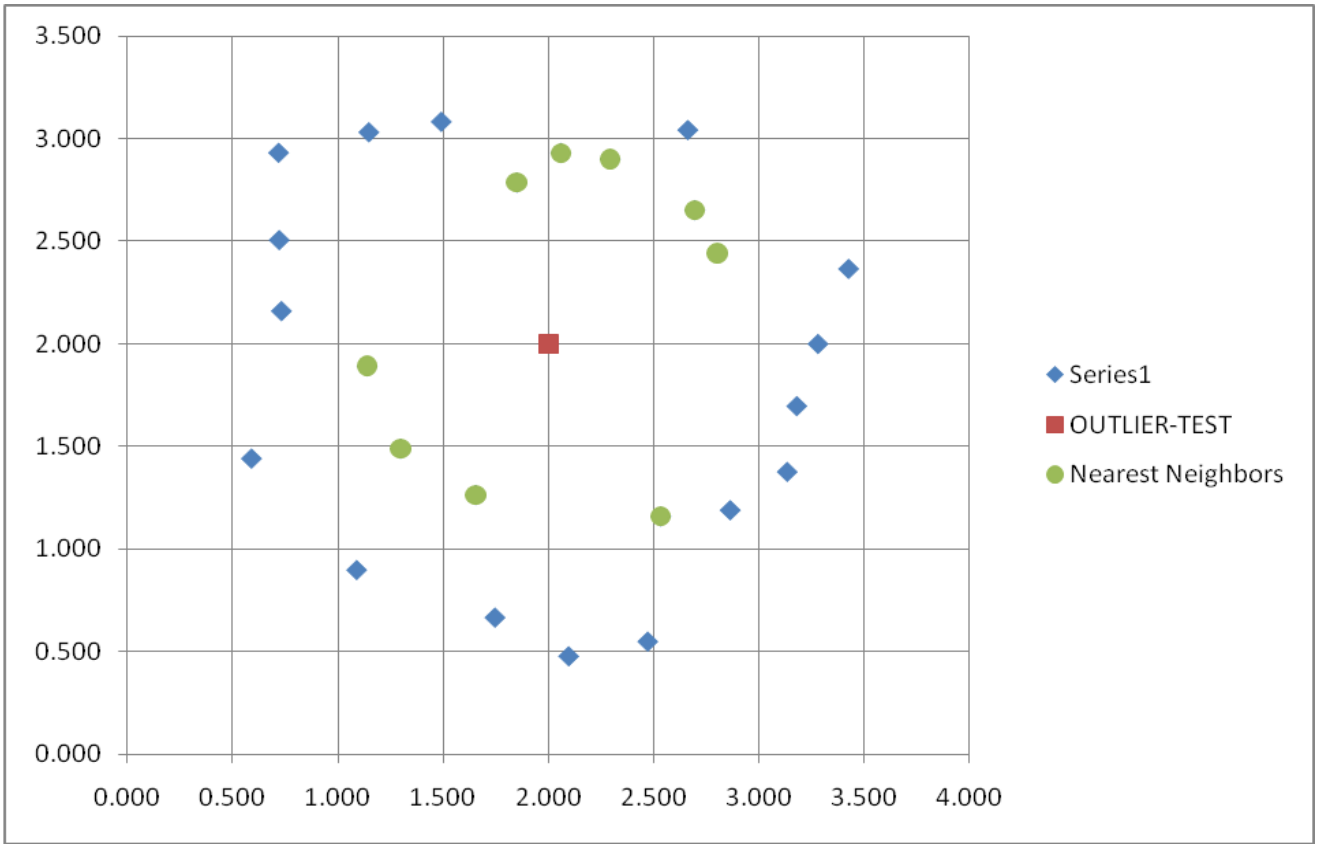

Figure 9c – Experiment C-HO (see Section 6.3 and Table 1)

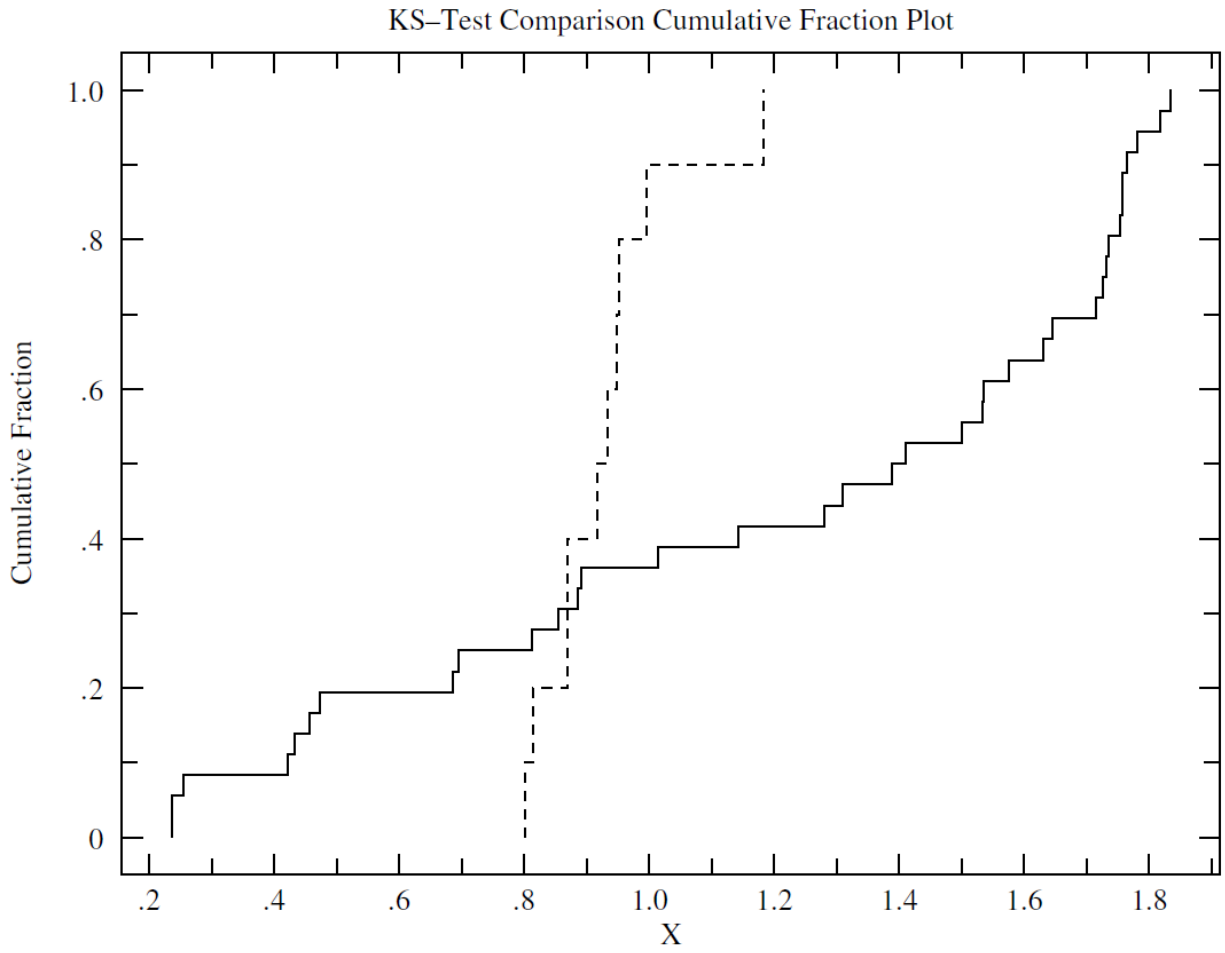

Figure 10 – Cumulative distribution plots for Experiment C-HO (see Section 6.3) used in the KS-Test to test if the two distance distribution functions $f_K(d,O)$ and $f_K(d,K)$ are drawn from the same population (Section 5). The ordinate X refers to the distance between data points in the sample. The solid line is the cumulative distribution of distances exclusively among the K nearest neighbors to the test data point. The dashed line is the cumulative distribution of distances between the data point and its K nearest neighbors (illustrated in Figure 9c).