Local outlier detection in data forensics:
data mining approach to flag unusual schools

Mayuko Simon

Data Recognition Corporation

Paper presented at the 2012 Conference on Statistical Detection of Potential Test Fraud in Lawrence, KS. This paper is revised version.

All correspondence and citation permission should be directed to Mayuko Simon, Data Recognition Corporation - 13490 Bass Lake Road, Maple Grove, MN 55311 (msimon@datarecognitioncorp.com).

THE WORK DOES NOT REPRESENT THOUGHTS OR OPINIONS OF COMPANY.

Introduction

Due to the high-stakes nature of the state standardized assessments, it is prudent for a governing body to ensure that the results from the assessments are based on effective instruction and true student achievement. While there are many ways to achieve meaningful acquisition of student knowledge via test scores, there are also ways to obtain higher test scores that are less related to actual learning. Some school administrators and/or teachers are thought to feel increasingly motivated to help students during test administration or change students' responses thereafter. This situation creates the need for data analysis to identify this aberrant behavior. Such analyses are collectively referred to as *data forensic*s. U.S. Secretary of Education Arne Duncan issued a policy letter (dated June 24, 2011) that urges states to "make assessment security a high priority" and "ensure that assessment development contracts include support for activities related to test security, including forensic analysis.", it is recommended by the Association of Test Publishers and the Council of Chief State School Officers (2010) that rules and procedures be adopted that respond to instances of test administration irregularities.

Most pertinent research methods used in the educational disciplines are rooted in statistics. These statistical methods include a wide array of techniques from the simplest univariate methods to sophisticated model-based multivariate and nonparametric techniques. One of the simplest methods is a univariate distributional technique focusing on a single metric such as the wrong-to-right erasure count average. It flags schools that exhibit extremely high average erasure counts relative to the state average in terms of the standard deviation. On the other end of the spectrum are the model-based multivariate techniques often based on regression analysis. In this scenario, a *dependent variable* (e.g. this year's test score) is modeled based on one or more *independent variables* (e.g. last year's score or average wrong-to-right erasure count). A school is flagged if the observed dependent variable differs significantly from the model's prediction. As different as these statistical outlier techniques are, they share a common trait: the schools they flag as outliers are different from *all* other schools. For this reason, we refer to these outlying schools as *global* outliers; they are outliers with respect to all other data points.

Schools with suspicious behavior may not display sufficient extremity to make them outliers in comparison to *all* schools. Nevertheless, it is reasonable to assume that their scores will be higher than that of their *peers*—schools that are very similar in many relevant aspects. Observations that are extreme with respect to their peers but not necessarily with respect to the entire data set are *local* outliers. For example, a rectangle and a triangle in Figure 1 can be considered local outliers. Those points are within the overall data range, but lie outside the two obvious clusters.

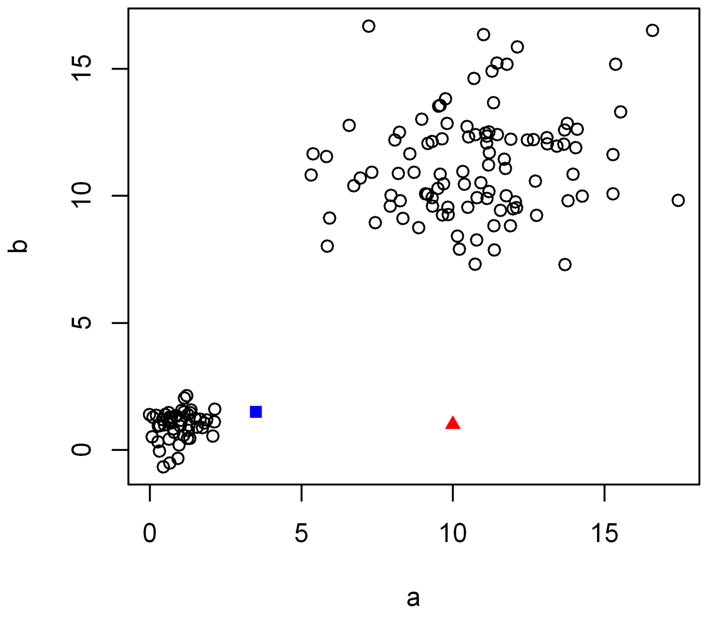


Figure 1. Illustration of local outliers (adapted from Breuning, et. al (2000)).

Traditional statistical data forensic techniques lack the ability to detect local outliers, hence we turn to data mining. Data mining has been applied in a great number of fields, including the education discipline (e.g., Baker, Gowda, & Corbett, 2011; Bravo & Ortigosa, 2009; Pardos & Heffernan, 2009).

Even in data mining, most outlier detection algorithms are global (Chandola, 2007). The prototypical local outlier detection algorithm is the Local Outlier Factor (LOF) by Breunig, Kriegel, Ng, and Sander (2000). LOF has been successfully employed in a number of applications, including computer network security (Lazarevic, 2003) and anomaly detection in the financial industry (Chen, 2007). To the best of our knowledge, our current work represents the first application of local outlier detection for cheating in the context of educational test scores. This is the first contribution of this work.

LOF in its original form is inadequate for our purpose. Consider a school *s* and its peers—the schools most similar to *s*. If *s* was moderately different from its peers in most respects, as a cumulative effect, *s* could become substantially different overall, and LOF could flag it. Being moderately dissimilar in most respects is not necessarily indicative of cheating. We wish to find schools that are very similar to the peers in most respects (in terms of most independent variables) but differ significantly in this year's score (the dependent variable). Since this school is similar to its peers in most independent variables, it is reasonable to assume that it should be similar in the dependent variable, as well. A significantly increased value of the dependent variable raises suspicion. Towards identifying such schools, our second contribution is to integrate model-based outlier detection with local outlier detection through incorporating the concepts of dependent and independent variables into local outlier detection.

In the proposed study, we employ a method with origins in data mining to identify schools that have outlying scores relative to their peer schools. We call this method Regression based Local Outlier Detection algorithm (RegLOD). For each school, the proposed method first identifies the school's peers—the most similar schools in terms of the relevant independent variables. RegLOD then draws the empirical distribution of the dependent variable from the peer schools and determines the empirical p-value of the school under investigation being an outlier with respect to its peer schools.

The key assumption underlying RegLOD is as follows. When most independent variables are very similar, we expect the dependent variable to be similar, as well. This assumption is very reasonable and is frequently exploited: this is the principle on which regression trees or nearest neighbor regression are built (Hastie, 2009).

We will demonstrate that local outlier detection is beneficial in detecting cheating and we will also show that RegLOD is capable of identifying local outlier schools which were not detected by the global outlier detection algorithms we routinely apply.

Method

Data Preprocessing

The data was obtained from a large scale standardized state assessment test and it contains the following variables: current and previous (2011 and 2010) years' scale scores aggregated to school level; the student cohort's score for the previous year (for grade 4, the test score when they were in grade 3 the previous year) aggregated to school level; average wrong-to-right erasure counts for current and previous years. All the scale score variables were transformed to logit scale (standardized ability score). The logits across different grades were not vertically scaled, thus logits across grades were not directly comparable. For example, logit of 2011 Mathematics are not on the same scale as 2010 cohort Mathematics. The wrong-to-right erasure counts were transformed into percentile score. For the purpose of detecting aberrant behavior with 2011 Reading, the dependent variable would be 2011 Reading logit score, and the independent variables would be 2011 Mathematics logit, 2010 Reading logit, etc.

Analysis Overview

The goal of the study is to detect local outlier schools—schools that are outliers with respect to their peers, not necessarily with respect to all schools. We used the proposed algorithm to this end.

First, we selected the independent variables to include in the analysis. This was done through multiple regression and assessing the explained variance of the model (the R-squared of the model) using all observations. We found all models (grade and subject combination) had high model R-squared, so all independent variables described above were included in the analysis.

Next, we applied RegLOD and the flagged schools were manually inspected. We describe RegLOD and this process in details in the subsequent sections.

RegLOD Algorithm Overview and Definitions

Algorithm 1: Main Steps of the RegLOD Algorithm.

1. Peer-group identification
    a. Compute the weight *w* of each independent variable for the distance calculation.
    b. For each school *s*, extract its *peer* schools. A school is a peer of *s*, if its weighted Euclidean distance to *s* (with respect to weights *w* from Step 1) are less than *Dist*—a user-supplied parameter.
2. Local outlier detection within the peer-groups

    For each school *s*,

    a. Draw the empirical distribution of the dependent variable from its peer group (excluding *s* itself) through bootstrap sampling with replication.
    b. Obtain the empirical one-tail (two-tail if all extremities should be flagged) p-value for *s* by drawing the empirical school mean-score distribution for the group of schools.
    c. Obtain the average empirical p-value over the bootstrap sampling.

    Flag schools with small p-values (<= 0.05) or schools with a low number of peer schools (10 or less).

The defining characteristic of the proposed method is to detect outliers locally. A school *s* is called an **outlier** if its score (dependent variable) is extreme with respect to that of its *peer* schools, namely the schools that are similar in relevant aspects (independent variables). Formally, a school *j* is a **peer school** of *s* if the weighted Euclidean distance between *s* and *j* in the feature space span by the independent variables is less then *Dist*, a user-supplied parameter. Technically, the Euclidean distance is a measure of *dis*similarity, but we will use the terms `distance' and `similarity' interchangeably. The set of schools that are the peer schools of *s* are called the **peer group**.

In the definition of a peer school, the term `relevant' alludes to the fact that not all variables are equally important. Therefore, we need to assess the relative importance of each independent variable and weigh them accordingly in the weighted Euclidean distance calculation.

As Algorithm 1 shows, the RegLOD algorithm has two key steps. First, we assess the importance of the variables, which in turn informs the identification of the peer group for each school *s.* Second, once the peer group is identified, we determine whether *s* is an outlier with respect to its peer schools. In the subsequent sections, we look at these two steps in details.

Assessing the Importance of Independent Variables and Identifying Peer Schools

Algorithm 2. Computing the weights used in distance calculation

1. Initialize all weights to 1/*K*, where *K* is the number of independent variables.
2. Compute the weighted Euclidean distance between school *i* and *j* as follows.

$$D_{ij} = \sum_{k=1}^{k=K} w_k \left(x_{ki} - x_{kj}\right)^2 , \qquad (1)$$

where D_{ij} is the distance between school *i* and *j*, w_k is the weight for *k*th independent variable and x_{ki} is the independent variable *k*'s value for school *i*.

3. For each of the school *s*, form its current peer group *P* selecting closest pre-determined number of schools.
4. Perform regression analysis and obtain the coefficients. Normalize the coefficients such that the absolute values of the coefficients sum to 1. Let c_{sk} denote the coefficient of the *k*-th independent variable for school s.
5. From steps 4 and 5, we obtained multiple sets of transformed coefficients, namely one set of coefficients for each school. The weight w_k for the k-th independent variable is the mean of the corresponding coefficients

$$w_k = \frac{1}{S}\sum_{s=1}^{S} c_{sk}, \tag{2}$$

   where *S* is the total number of schools.
6. Repeat steps from 2 to 5 until the sum of square of the coefficients difference between iterations are less than a certain threshold.

We assess the importance of the independent variables and identify peer schools simultaneously through an iterative algorithm shown in Algorithm 2. The key idea is to assess the importance of the independent variables through their capability of predicting the dependent variable within the peer group. To this end, we use multiple regression and the weight of each variable used in the distance calculation is proportional to the corresponding regression coefficient.

Specifically, for each school *s*, we initially set the weights *w* of all variables to 1/*K*, where *K* is the number of independent variables. Then the weighted Euclidean distance to all other schools are computed (Equation 1) and the peer schools are identified. The peer schools for weights calculation are pre-determined number of the schools that are closest distance from *s,* (in this paper, 199 was used). Next, a regression model is built using the peer schools, predicting the dependent variable. At the end of this iteration, the output is a set of regression coefficients for each school *s*. We then update the weight w_k of the k-th independent variable as the average of the corresponding coefficient (Equation 2).

Updating the weights used in the distance calculation naturally changes the distances between the schools and consequently changes the peer groups for each school. We iterate through steps 2 to 5 updating the weights and re-identifying the peer schools until convergence. Convergence is attained when the sum of squared coefficients change less than a small positive number ε (ε=.001 was used in this paper).

Implementation details and parameter selection. In a study with a large number of schools, re-computing the all-pair distances to determine the peer schools may represent an unreasonable computational burden. In such cases, the distance weights can be computed using a relatively small sample of the schools. In our current study, we selected 100 schools randomly to compute the weights for distance calculation.

The value for *Dist* in the current analysis was 0.03. The value was chosen arbitrary by observing how many peer schools can be formed for each school. When the value is too small, the number of peer schools would be too small and too many schools will be flagged as outliers since many schools cannot find enough counts of peer schools. Conversely, when the value is too large, e.g., half of the population may be chosen as peer schools for a school, the peer school may contain schools that are not really peers to the specific school.

Identifying outliers in the peer group

Outlierness of the test score (dependent variable) of the specific school relative to its peer schools are assessed through the empirical p-value. We applied bootstrapping to draw the empirical distribution of the logit score within the peer group and calculated the empirical p-value for the observed logit score.

Final list of the flagged schools were inspected manually. The final manual inspection also considered current and previous years' test level wrong-to-right erasure percentile scores. The wrong-to-right erasure percentile scores were not included in the process to group peer schools since the peer schools may not necessarily have similar erasure percentile scores.

Evaluation Method

The results of the RegLOD were compared to the results (Outlier Score) of other data forensic methods described in Plackner (2012). These methods include: scale score analysis (SS) where two year's scale score change of the same subjects were examined; performance level analysis (PL) where two year's proportions of proficient and above were compared; Regression (Reg) analysis where scale score of one subject was predicted by scale score of other subject; Rasch (Rasch) analysis where residual of student level analysis was conducted; and cohort scale score (SSco) analysis where cohort scale score of two years were examined. For more detail explanation of each analysis, refer to Plackner (2012). These methods explained in Plackner (2012) are the analyses based on statistical methods and the flagged schools are global outliers. The standardized residual from multiple regression using all observation was also used in the comparison.

Results

We devote the first half of this section to showing the descriptive statistics for the RegLOD, and in the later half will demonstrate on the real-world example of grade 4 Reading that local outlier detection and RegLOD in particular is valuable tool in data forensics as it was capable of identifying schools with suspicious behavior that the more standard techniques missed.

Descriptive Statistics

All model R-squared were very high (Table 1), verifying that the independent variables used in the analysis are indeed predicting the dependent variables and appropriate to use in the distance computation. The model R-squares were above 0.88 for Mathematics and Reading for grades 4 to 8. Through the multiple regression analysis, all variables were found to be significant predictors in all grades and subject combination, except grade 8 Reading, which cohort Mathematics was not a significant predictor, and grade 7 Mathematics, which the cohort Reading was not a significant predictor.

Table 1. The R-square of the multiple regression model using all observations

Subject	Grade	R-squared
R	4	0.91
R	5	0.90
R	6	0.93
R	7	0.92
R	8	0.94
M	4	0.88
M	5	0.89
M	6	0.89
M	7	0.91
M	8	0.92

Figure2 shows the distributions of logit scores for the variables used for grade 4 Reading. The variables were mostly slightly negatively skewed and ranged between -1 to around 3. These logits were not vertically scaled, so the comparison of logits scores across different grade should not be made.

Table 2 shows the weights used in the weighted pair-wise Euclidean distance calculation. The largest weights in Mathematics were 2011 logit score for Reading in all grades. For Reading, a variable with largest weights differed depending on grades. With grades 4 and 5, the 2011 Mathematics (P1) had the largest weight. With grades 6 and 8, the cohort's Reading had the largest weights. Grade 7 had largest weights with 2010 Reading. The cohort logit score of another subject (i.e., Reading for Mathematics model, Mathematics for Reading model) seem to have very small weights regardless subjects and grades.

Figure 2. Variables used in Grade 4 Reading analysis with all observation (1603 schools).

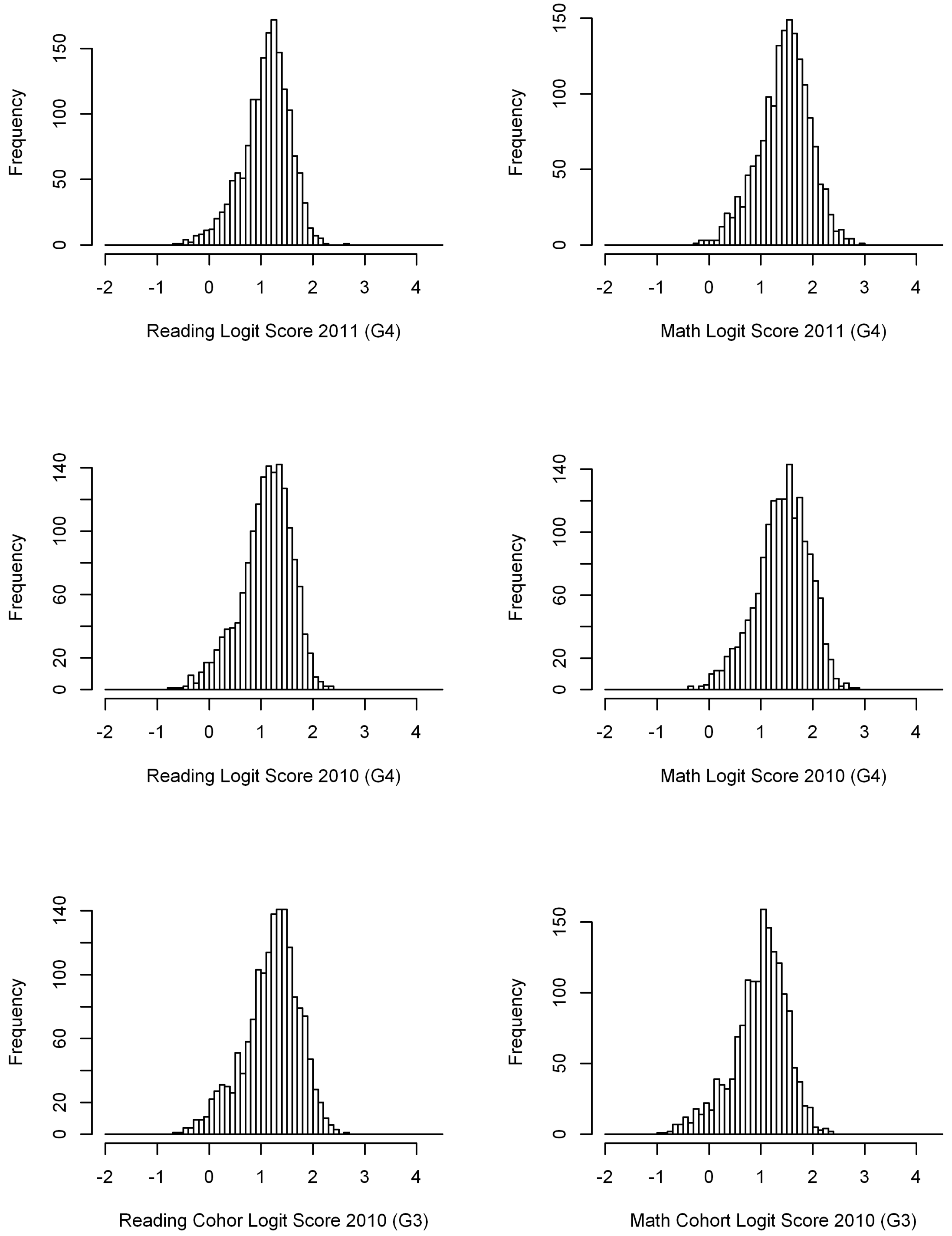


Table 2. Weights used in the Euclidean distance calculation to from peer group.

		2011 Logit Score	2010 Logit Score		2010 Cohort Logit Score	
		M or R	R	M	R	M
Subject	Grade	P1	P2	P3	P4	P5
R	4	0.37454	0.21180	-0.11270	0.23794	-0.06301
R	5	0.30482	0.22893	-0.11496	0.25870	-0.09259
R	6	0.27619	0.24746	-0.12622	0.30431	-0.03427
R	7	0.24754	0.26632	-0.12389	0.28320	-0.07790
R	8	0.20851	0.26684	-0.10840	0.38520	-0.01790
M	4	0.42298	-0.15515	0.24383	-0.07951	0.09744
M	5	0.36514	-0.13519	0.20929	-0.09919	0.19083
M	6	0.35133	-0.18089	0.25044	-0.09666	0.12017
M	7	0.33369	-0.18963	0.28817	-0.01410	0.13175
M	8	0.30469	-0.16628	0.24992	-0.04278	0.22135

P1: logit of another subject in the same year (e.g. for the model of Reading Grade 4, it is the logit score of Mathematics for grade 4 in 2011).
P2: Logit score of previous year's same grade Reading (e.g. for the model of Reading Grade 4, it is the logit score of reading for grade 4 in 2010).
P3: Logit score of previous year's same grade Mathematics.
P4: Logit score of Reading with cohort students (e.g. for the model of Reading Grade 4, it is the logit score of reading for grade 3 in 2010 including only cohort students).
P5: Logit score of Mathematics with cohort students.

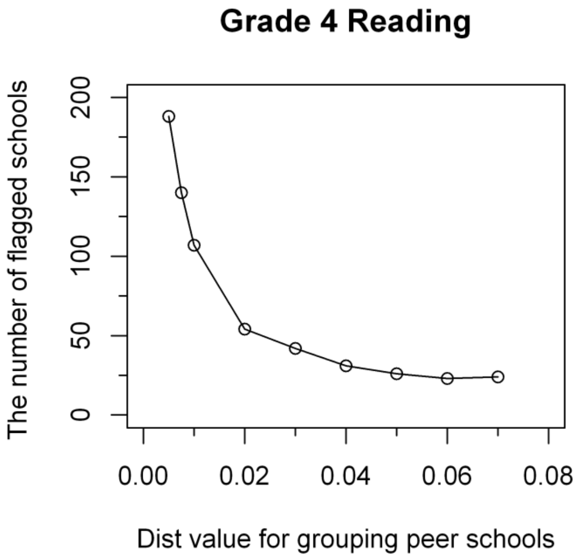
With *Dist* value 0.03, the proportion of flagged school ranged from 2.06 percent to 4.43 percent across all subjects and grades (Table 3). The *Dist* value of 0.03 was used in this paper to flag rather small proportion of schools. Different *Dist* value can be used, but will yield different proportion of flagged schools. When the *Dist* value was set larger, such as 0.05, the number of flagged school reduced but some schools had as many as almost half of the total number of schools as peer schools and the size was considered too large to detect local outliers. When the *Dist* value was reduced to 0.01, the number of peer schools for each school reduced, but in return, it increased the number of schools that could not find more than 10 peer schools and those were flagged for not finding enough peer schools. For example, for grade 4 Reading, the *Dist* value of 0.01 flagged more than 100 schools and many of them had less than 10 peer schools. Examining the flagged school with the *Dist* values of 0.01, many schools did not have clear indication of aberrant behavior, thus the *Dist* value 0.01 was considered too small. Figure 3 shows the relationship between the *Dist* value and the number of flagged school for grade 4

Reading. Although results are not shown in this paper, most schools that were flagged when the cut-off value was large were also flagged with the small *Dist* value.

Table 3. The number of schools used in the analysis and the number of schools flagged.

			Local weights (Dist = 0.03)	
Subject	**Grade**	**Total School N**	**Flagged School N**	**Proportion Flagged**
R	4	1603	42	2.62
R	5	1493	38	2.55
R	6	1056	46	4.36
R	7	836	21	2.51
R	8	836	22	2.63
M	4	1603	32	2.00
M	5	1493	63	4.22
M	6	1056	37	3.50
M	7	836	38	4.55
M	8	836	36	4.31

Figure 3. The relationship between the Dist value and the number of flagged schools.



Grade 4 Reading Results

In grade 4 Reading, 1603 schools were examined and among them, 49 schools were flagged and shown in Table 4. The analysis used logit scores, but since logit scores were not vertically scaled, it is not possible to compare between different grades' logit scores. For example, 2010 cohort's Reading logit score in grade 3 cannot be compared to 2011 grade 4 Reading logit score. Thus, Table 4 shows the percentile rank of the average scale score instead of the logit scores used in the analysis. Percentile rank gives relative standing of the school compared to other schools, and gives better sense of understanding with aberrant behavior. Table 4 also shows the percentile rank of wrong-to-right erasure counts in 2010 and 2011, and outlier score from other methods described in Plackner (2012). The standardized residual from multiple regression using all schools were also obtained and shown in the table. The outlier scores from Plackner (2012) and the standard residual are the results of data forensic analyses detecting global outliers. The outlier score exceeding 10 were considered outlier schools for the methods in Plackner (2012). With standardized residual, exceeding 2 would be considered as outliers. Table 4 shows that local outlier sometimes overlap with global outliers, but not all the time. All schools need to be investigated further.

The flagged schools were manually examined. The current and previous years' percentile scores and erasure percentile score were checked to ensure that the results are sensible. Having percentile score allowed us to examine if the results made sense. Although not all schools with aberrant behavior will have high erasure percentile score, but having relatively high erasure percentile score gives more doubt. Three groups of flagged schools emerged: 1) Schools that experienced a large percentile score increase in the cohort accompanied with high erasure percentile rank (e.g., erasure percentile rank > 60); 2) Schools that do not have very high erasure percentile rank, but the percentile rank of the scores had a large increase; 3) Schools that have high achieving students giving the schools extremely high percentile score in current and previous years, or opposite (low achieving schools); 4) Schools that do not have a clear increase in percentile rank score or higher erasure percentile rank. Among the four groups of schools, groups 1 and 2 exhibit apparently suspect behavior. Group 3 schools seem to be just extremely high/low achieving schools and their percentile scores are consistent across years. We hypothesize that they were flagged because their extremely high or low achievements make them appear aberrant even in comparison to their peers or simply their achievement is too extreme or cannot find enough peer schools. Group 4 schools—although extreme in comparison to their peer schools--show no suspicious behavior to the extent we can ascertain from our data. With 42 schools flagged with grade 4 Reading, 18, 10, 11, and 3 schools seems to belong group 1, 2, 3, and 4, respectively. Most of the schools in group 3 were flagged because the schools did not have more than 10 peer schools.

The school number 1 in Table 4 was flagged because the peer N was 4, indicating this school's score pattern was very rare. The percentile score data shows that this school shows an evidence of aberrant behavior: it has 2010 cohort Reading 26 percentile and 2011 Reading is 95 percentile. This school was flagged with the cohort scale score analysis (SSCo), but not the other methods in Plackner (2012). The 2010 cohort Mathematics was also 20 percentile, so the cohort students were relatively lower achieving students compared to all schools in the states. However, this school (different cohort, students who are one year older) had improved their percentile score in 2011, grade 4 students in 2011 were 95 percentile and 52 percentile with Reading and Mathematics, respectively. The grade 4 students in 2010 in the same school were 100 percentile and 93 percentile for Reading and Mathematics, respectively. The RegLOD fond only four peer schools including the school, and the wrong-to-right erasure was 96 percentile in 2011, thus there are reasonable evidences that this school need further scrutiny.

Figures 4 to 6 show histogram plots of the variables used with the peer schools for three schools. The vertical lines indicate where the schools locate among the peers based on the weighted pair-wise Euclidean distance. The group of peer schools was formed around the specific school, so the vertical lines (the specific school) tend to sit around the center of the distributions, especially with the variable with large weights. With grade 4 Reading, 2011

Mathematics had largest weights followed by 2010 Reading and 2010 cohort Reading. Cohort Mathematics had very small weights.

Figure 4.a and 4.b show the school number 4 in Table 4. Figure 4.a is in logit scale which were used to group peer schools, and Figure 4.b. is in percentile scale for better understanding of the changes in school's score across years (e.g., 2010 cohort Reading to 2011 Reading) compared to all schools in the state. The spread of the distribution with percentile score is wider than logit score since logit score has a somewhat bell shape as opposed to uniform distribution of percentile when all school would have been plotted. The figures show only the peer schools for the school number 4.   This school was 23 percentile with 2010 cohort Reading and 2011 Reading was 76 percentile, which was quite large increase looking at the percentile. The wrong-to-right erasure percentile was 52, and it is not extremely large. An analysis using scale score increase across 2 years (SS and SSCo) or any other methods in Plackner (2012) (PL, Reg, Rasch, and WR) did not flag this school, but it was flagged by the RegLOD algorithm. The standardized residual from multiple regression was 2.91, so if we used multiple regression, this school would be flagged. Therefore, this school was a local outlier through RegLOD and also a global outlier when we used multiple regression method. Histogram distributions show how different this school is compared to the peers. Looking at Figure 3 histograms, School number 4's 2011 Mathematics was in right tail (higher score relative to the peer schools), but 2010 Reading percentile and 2010 Reading cohort percentile for the school was in slightly lower than center of the distribution. However, the school had higher percentile score than any other peer schools with 2011 Reading. The school was well within the distribution with independent variables compared to the peers, but the dependent variable (2011 Reading) was an outlier in comparison to the peer schools. This is an evidence of local outlier.

Figure 5 shows school number 11 in Table 4 in percentile score. This is an example of local outlier school with relatively low achievement. This school's 2010 cohort Reading was 13 percentile and 2011 Reading was 42 percentile. The wrong-to-right erasure was 96 percentile. With high erasure percentile and large increase in percentile score (relative standing in all population) gives an indication of aberrant behavior of this school. However, this school was not flagged by methods in Plackner (2012), but the standardized residual from multiple regression was exceeding 4, thus very extreme from multiple regression perspective. This school is a local outlier and a global outlier when multiple regression was used. The histograms in Figure 4 show that the school is in the center of the distributions with 2011 Mathematics, 2010 Reading, and 2010 cohort Reading, but the school was an outlier in 2011 Reading distribution. The school's 2010 cohort Mathematics was also very high compared to the peers. However, this was probably due to the very small weight for Mathematics cohort in computation of weighted paired Euclidean distance.

Figure 6 shows school number 20 in Table 4. This is an example of high achieving school. This school was 77 percentile with 2010 cohort Reading and 95 percentile with 2011 Reading. The erasure percentile was 87, so it is relatively high, but not extreme. This school was not flagged by methods in Plackner (2012) or multiple regression, but flagged as a local outlier by RegLOD. Figure 5 shows that the peer schools for this school are higher achieving schools than Figure 4. This school was in the center of distribution with 2010 cohort Reading, and 2010 Reading, however, this school was higher than majority of peer schools in 2011 Reading and the empirical p-value was 0.026.