Figure 3.a. An example of flagged schools (School number 4 in the Table 4) with logit sca

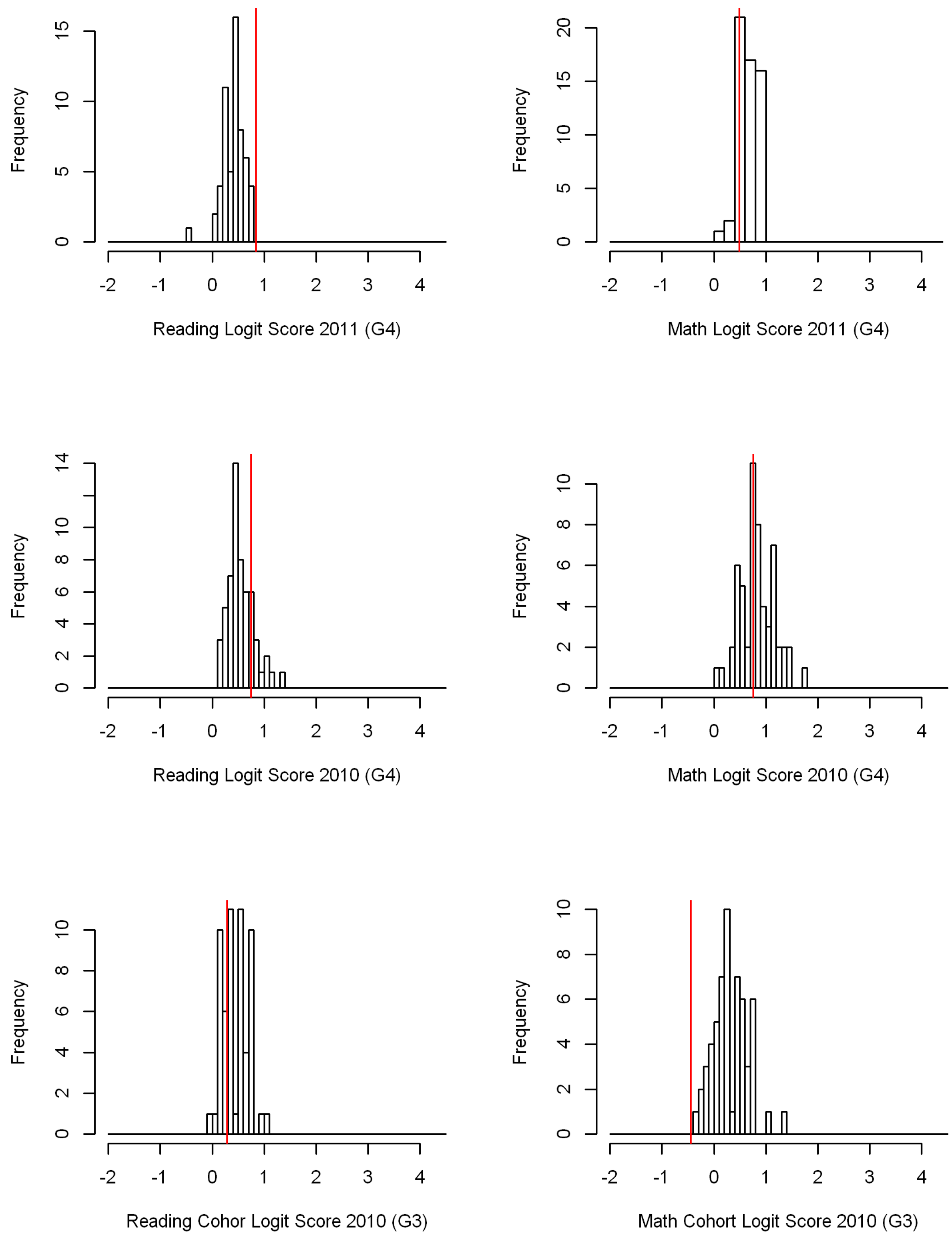


Figure 3.b. An example of flagged schools (School number 4 in the Table 4) with percentile scores.

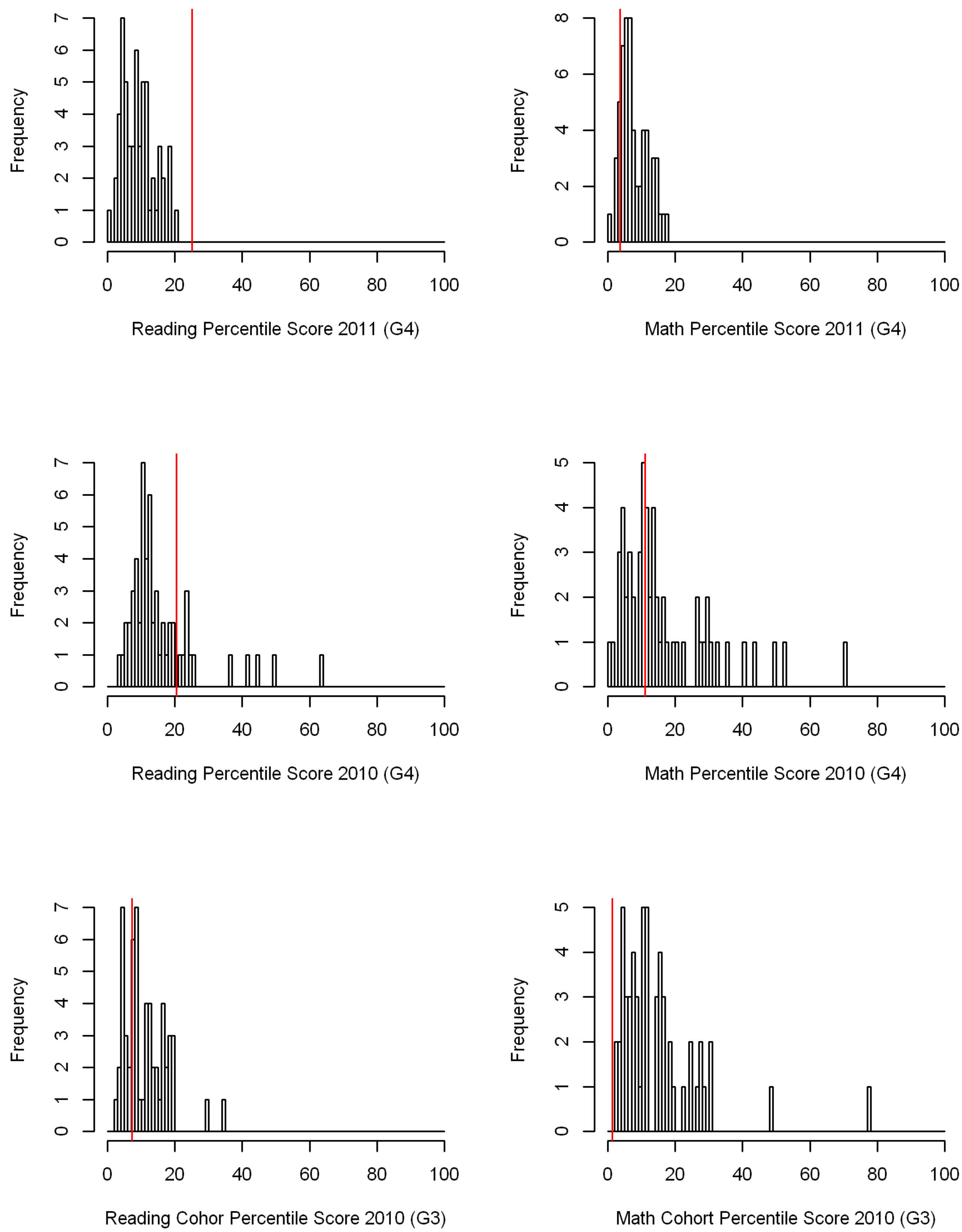


Figure 4. An example of flagged schools (School number 11 in Table 4) with percentile scores.

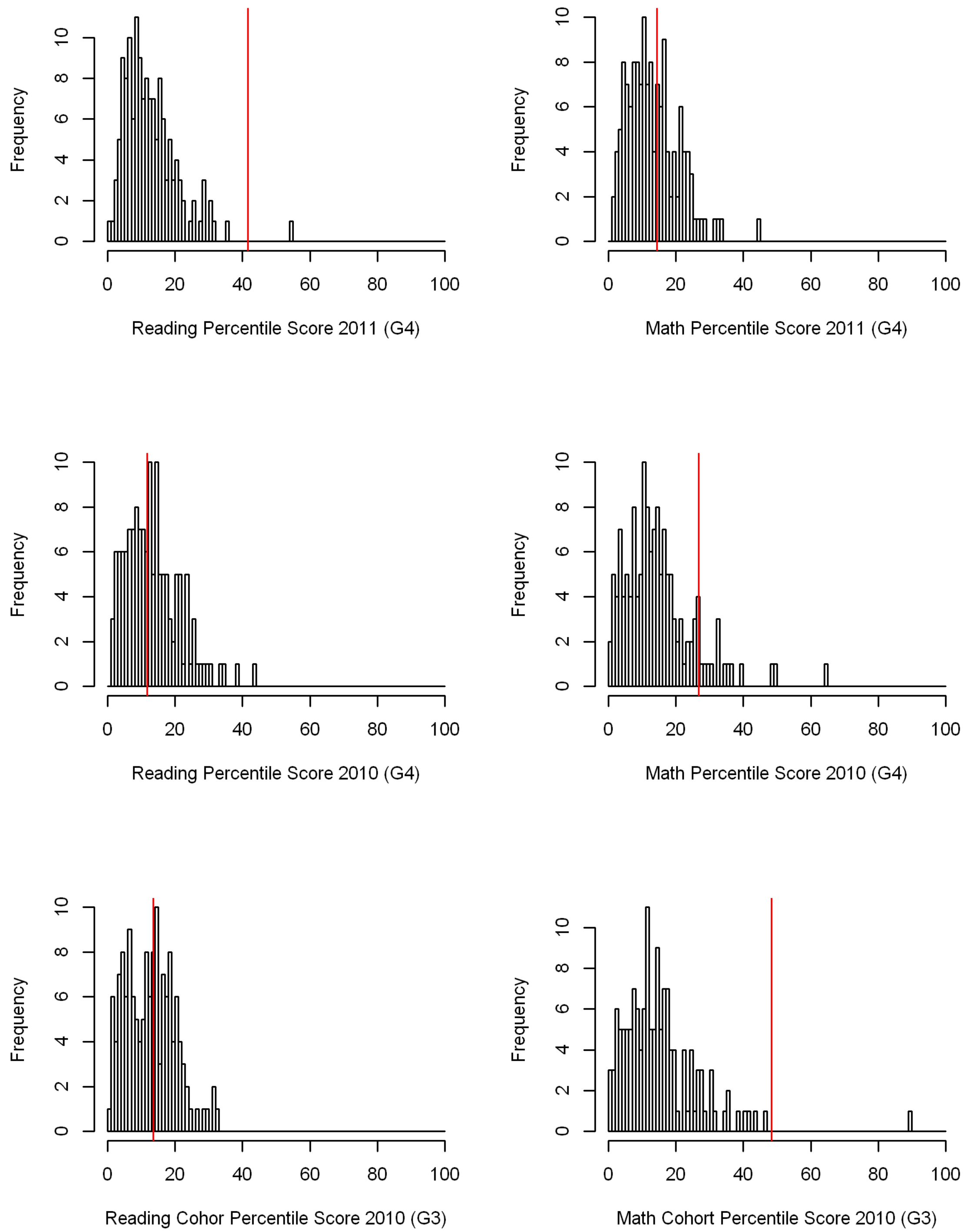
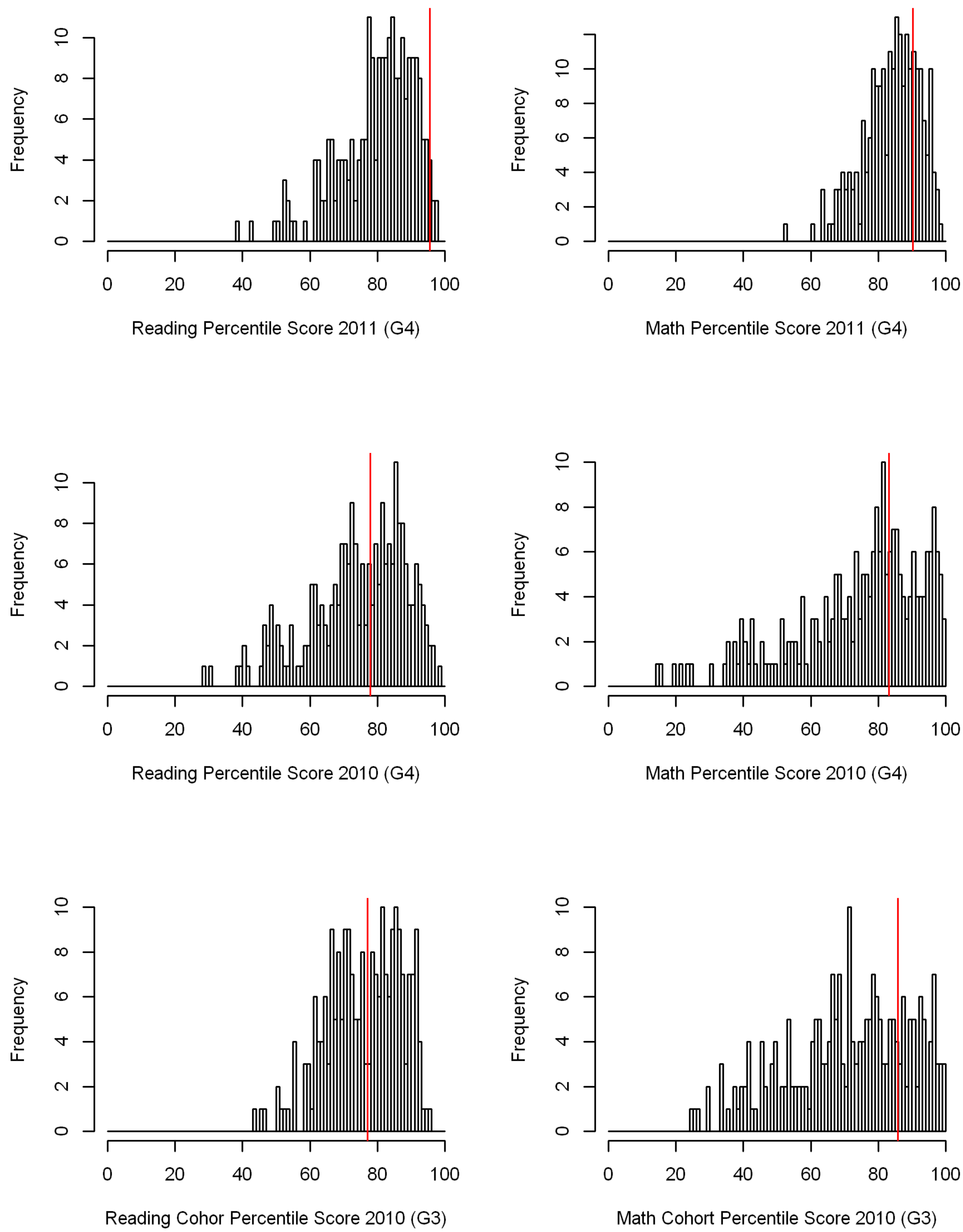


Figure 5. An example of flagged schools (School number 20 in Table 4) with percentile scores.



Discussion and Conclusion

In this paper, we introduced a novel algorithm, RegLOD, to detect local outliers in data forensics. We demonstrated the algorithm on real data from a large standardized test and compared it to the other—global—methods of outlier detection. Local outlier detection and specifically our proposed methodology shows great promise in data forensics. A great strength of the method is its increased sensitivity: RegLOD can flag schools that were missed by methods aimed at detecting global outliers. While it flagged 11 schools in grade 4 Reading possibly incorrectly solely for their extremely high/low achievement, by and large, it flagged schools that indeed exhibited suspicious behavior. It flagged some schools that were flagged by different statistical methods, as well, but it also correctly flagged some schools that were not detected by competing methods. Manual inspection of these schools' data revealed suspicious behavior.

The critical part of the RegLOD methodology is the identification of the peer schools. In this paper, the groups were formed using regression weights and *Dist* value of 0.03. Different methodology can be used in the distance computation and different *Dist* value can be used. With large *Dist* value, it creates large peer groups, and large proportion of schools in a peer school loses the benefit of local outlier detection algorithm, so the *Dist* should not be too large. However, a *Dist* value that is too small will not have enough peer schools and some schools could be flagged as local outlier when it should not.

The variable used in this analysis included cohort score from the previous year. Therefore, the grades that can be analyzed with RegLOD in this example is limited. For example, the analysis with grade 11 was not possible for this data set since grade 10 students were not given an assessment. Same is true for grade 3. However, the same approach of RedLOD could be used with different set of predictors to detect local outliers.

The analysis in this paper concentrated only on variables that were a priori known to be important. We also performed variable pre-selection to ensure that all variables brought forward into the RegLOD analysis were relevant. This approach was primarily guided by interpretability rather than the method's capability of handling irrelevant variables. Data mining, where this method finds its roots, is often used as an exploratory technique with a large number of predictors of questionable relevance to the dependent variable. The distance metric learning aspect of RegLOD (where the weights for the (dis)similarity metric are `learned' from the data through multiple regression) is aimed at this precise problem. When irrelevant variables are present, their regression weights will be small and hence these variables will have very limited influence on the peer group selection.

RegLOD have shown great promise in data forensic and it is a valuable addition to our data forensic tools. Its applicability is not limited to cheating detection in educational testing. Given its robust design, specifically its model-based design (the concept of dependent and independent variables) and its ability to adapt (distance metric learning) makes it applicable to a wide range of outlier detection problems. We continue to study its capabilities, extend and apply it to other contexts and tasks.

Reference

Baker, R.S.J.d., Gowda, S., Corbett, A.T. (2011). Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. Proceedings of the 4th International Conference on Educational Data Mining, 179-188.

Bravo, J. & Ortigosa, A. (2009). Detecting Symptoms of Low Performance Using Production Rules.Proceedings of the 2nd International Conference on Educational Data Mining, 31-40.

Breunig, M. M., Kriegel, H., Ng, R. & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. Proceedings of the ACM SIGMOD International Conference on Management of Data, Dalles, TX.

Chandola, V., Banerjee, A., & Kumar, V. (2007). Anomaly Detection: A Survey. Technical Report, University of Minnesota

Chen, M.-C., Wang, R., & Chen. A.-P. (2007). An Empirical Study for the Detection of Corporate Financial Anomaly Using Outlier Mining Techniques. In ICCIT '07: Proceedings of the International Conference on Convergence Information Technology, pages 612–617.

Hastie, T., Tibshirani, R., & Friedman., J. (2009) Elements of statistical learning. Second edition. Springer.

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., Srivastava., J. (2003) A Comparative study of anomaly detection schemes in network intrusion detection. Third SIAM Conference on Data Mining, San Francisco.

Pardos, Z. A., & N. Th. Heffernan (2009). Determining the Significance of Item Order in Randomized Problem Sets. Proceedings of the 2nd International Conference on Educational Data Mining, 111-119.

Plackner, C. (2012). Data forensics: A compare and contrast analysis of multiple methods. Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.