Measuring the Power of Learning.®

Research Report
ETS RR–17-15

Exploring Online Learning Data Using Fractal Dimensions

Hongwen Guo

April 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Exploring Online Learning Data Using Fractal Dimensions

Hongwen Guo

Educational Testing Service, Princeton, NJ

Data collected from online learning and tutoring systems for individual students showed strong autocorrelation or dependence because of content connection, knowledge-based dependency, or persistence of learning behavior. When the response data show little dependence or negative autocorrelations for individual students, it is suspected that students are randomly guessing the answers or that they are inconsistent in learning behavior. In addition, the global and local rates of correct responses may reflect students' proficiency in the learning process. This study shows that the dependence of online data may be characterized by the fractal dimension as a summary statistic locally and globally. The rate of correct responses and the global and local fractal dimensions of individual students' responses may indicate their learning behavior in short and long learning windows. The results may shed light on when individual students are experiencing difficulties in the learning process.

Keywords Response accuracy; dependence; fractal dimension; self-similar processes

doi:10.1002/ets2.12143

With advances in technologies, more and more schools and teachers have been using online teaching, learning, tutoring, and homework systems to assist student learning. For example, there were 841,687 registrations for online courses for HarvardX and MITx (Ho et al., 2014) from fall 2012 to summer 2013, and millions of K–12 students use online learning and homework systems. The online platforms track students' every click as they use instructional resources, complete assessments, and engage in social interactions. These data have the potential to help researchers identify, at a finer resolution than ever before, what contributes to students' learning and what hampers their success. Big data collected online may provide additional value in understanding student learning strategies and helping teachers in instruction, which is particularly useful for K–12 education.

As Fayyad, Piatetsky-Shapiro, and Smyth (1996) have pointed out, one basic problem is mapping raw data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact, more abstract, or more useful (e.g., a summary index, a predictive model). Commonly used summary statistics such as means and standard deviations may not be meaningful enough to describe the online learning processes. In this study, time-series techniques (including more suitable summary statistics, such as the dependence index) were sought to model students' response processes instead of final scores.

Unlike responses to educational assessment and tests, where students have limited time to respond to limited test questions/items (where most likely the number of items is in the double digits), online learning and homework solving usually do not have a time limit (instead it may last for one school year, for example), and it is not unusual to collect thousands of responses and clicks to homework questions for an online course during a school year for an individual student. In addition, students' proficiency levels may change in the process, and knowledge and skills change too. Therefore the traditional item response models based on item response theory (IRT) with and without response time (Lord & Novick, 1968; van der Linden, 2007) to model educational assessment may not be suitable for big online data. Data collected from online learning and tutoring systems for an individual student are expected to show strong and positive correlation or dependence because of content connection, knowledge-based dependency, or persistence of learning behavior. Because of violation of the local independence, this adds to another reason for not using IRT models for online data.

When the response data show little dependence or negative correlations for a student, it is suspected that students may be randomly guessing answers or that they may be inconsistent in learning behavior—a sign of problematic learning. Treating the responses as time-series sequences, Warnakulasooriya and Galen (2012) found that students' responses exhibit random walk- or Brownian motion-like characteristics. A random walk-like response exhibits irregularities or

Corresponding author: H. Guo, E-mail: hguo@ets.org sss

fluctuations about the expectations. Such a characteristic may be quantified by stochastic processes with or without dependence structures, such as fractional Brownian motions (fBms) and their fractal dimensions.

The fractal dimension of a fBm is the statistical index that characterizes the scaling property of self-similarity as well as covariance functions of the process (see Appendix A for details). A fractal analysis is used to evaluate and calculate autocorrelation and dependency between observations using various techniques (as shown in the Data and Fractal Analysis section) for long time-series sequences. It is particularly useful to simplify a dependency structure by using a few indices instead of many parameters.

The purpose of this study is to extract useful information from the voluminous data using stochastic processes. In this study, the long item response sequences from an individual student's learning process are treated as a time-series sequence. Because of the expected dependence among responses, this study attempts to characterize the learning process by a fractal dimension besides the rate of correct responses. I present statistical modeling of the time-series sequences by stochastic self-similar processes (Mandelbrot, 1983), such as the fBm. In this study, the fractal dimension of the stochastic process is used as a summary index to describe the overall dependency of the long time-series sequences and to characterize the persistence of a student's learning behavior. In addition, the confidence band of the estimated fractal dimension is provided to account for uncertainty.

More specifically, the study focuses on the following questions:

1. Is dependence observed between item responses?
2. Can the expected dependence structure be characterized by an overall fractal dimension index? If not, will the local fractal dimensions provide more information?
3. The rate of correct responses reflects a student's proficiency. Is it changing in the learning process?
4. Can analytical results on item response sequences alone provide useful information to detect whether a student is experiencing difficulty in the learning process?

The goal of such an analysis is to see whether combination of the fractal dimension estimation and the rate of correct responses could capture the moments when educators need to implement early intervention to improve teaching and learning before they obtain individual students' final exam or test scores.

This study is inspired by but different from Warnakulasooriya and Galen (2012) in the following ways: (a) The autocorrelation function (ACF) and other statistical tools are employed to explore whether a fractal dimension is plausible in the data sequences; (b) in addition to investigating the fractal dimension globally and locally, the global and local average rates of correct responses play an important role in interpretation; and (c) it is emphasized that the fractal dimension alone is not enough to classify a student. Instead, researchers need to pay attention to changes in fractal dimensions and average rates of being correct. More important, theoretical properties of fractal dimensions and calculation of these dimensions are described in Appendix A to clarify possible misunderstanding.

The remainder of the report is organized as follows. In the next section, stochastic self-similar processes and the fractal dimension are introduced. In Data and Fractal Analysis section, the data set collected from an online tutoring system is described. Analysis results are presented on item response accuracy. The fractal dimension is used to characterize the dependence in the response sequences, and the mean score (the rate of correct responses) sequences are examined to reflect students' learning curves. In the Discussion section, a summary and discussion are provided. Calculation of fractal dimensions is presented in Appendix A, and an exemplary scenario in the tutoring system is given in Appendix B.

Self-Similar Processes and Fractal Dimensions

Many complex geographical objects, such as earth terrain, weather, and DNA, are statistically self-similar, meaning that each small portion of the object can be considered a reduced-scale image of the whole. The degree of complication can be described by a quantity *D*, the fractal dimension, that has many properties of a dimension. The well-known self-similar processes are Brownian motion, fBms, Levy flights, and so on. Their fractal dimensions can be calculated mathematically. These processes are often used in modeling mathematical sequences, DNA sequences, financial data, traffic signals, texts, and so forth.

Fractional Brownian Motion

A real-valued stochastic process $Z = \{Z(t), t \in \mathbf{R}\}$ is self-similar with index $H > 0$ if, for any $a > 0$,

$$\{Z(at)\} =^d \left\{a^H Z(t)\right\},$$

where $=^d$ denotes the equality of the finite-dimensional distributions. This H is called the Hurst index of the process (Taqqu, 2003).

The process $Z = \{Z(t)\}_{t \in \mathbf{R}}$ has stationary increments if, for all $h \in \mathbf{R}$,

$$\{Z(t+h) - Z(h)\} =^d \{Z(t) - Z(0)\}.$$

A Gaussian self-similar with stationary increment (sssi) process $\{B_H(t)\}_{t \in \mathbf{R}}$ with $0 < H < 1$ is called a fBm. Its covariance function is

$$\mathrm{Cov}\,(Z(s), Z(t)) = \frac{\sigma^2}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right), \tag{1}$$

where $\sigma^2 = \mathrm{Var}\{Z(1)\}$. When $\sigma^2 = 1$, it can be shown that, for some constant C, as $t \to 0$,

$$E(Z(t+h) - Z(h))^2 \sim C|t|^{2H}. \tag{2}$$

It is shown mathematically that with probability 1, the Hausdorff dimension and box dimension (for definitions, see Appendix A) of the graph $\{t, B_H(t)\}_{0 \le t \le 1}$ are of the same value D, and $D = 2 - H$ (Taqqu, 2003). When $H = 1/2$, it is the regular Brownian motion.

The Hurst index and the fractal dimension, as well as the covariance structure in Equations 1 and 2, reflect the local topological structure of the process. Furthermore, fBm also has a global property that is characterized by H. Let $\{Z(t)\}$ be an H-sssi process, and let

$$X_k = Z(k+1) - Z(k), \quad k \in \mathbf{Z}.$$

The autocovariance function of the increment sequence $\{X_k\}$ is

$$\gamma(k) = EX_i X_{i+k} = \frac{\sigma^2}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right). \tag{3}$$

If $H \neq 1/2$, then

$$\gamma(k) \sim \sigma^2 H(2H-1)|k|^{2H-2}, \tag{4}$$

as $k \to \infty$.

In particular, for $k \neq 0$,

$$\gamma(k) = 0, \quad \text{if} \quad H = 1/2;$$

$$\gamma(k) < 0, \quad \text{if} \quad 0 < H < 1/2;$$

$$\gamma(k) > 0, \quad \text{if} \quad 1/2 < H < 1.$$

From the preceding equations one can show that, for $1/2 < H < 1$, the autocovariance tends to zero so slowly that $\sum_{k=-\infty}^{\infty} \gamma(k)$ diverges. In this case, one says that $\{X_k\}$ exhibits long-range dependence (or has long memory). For $0 < H < 1/2$, the autocovariance tends to zero quickly, and $\sum_{k=-\infty}^{\infty} \gamma(k) = 0$. In this case, it is said to exhibit antipersistence.

Computation of the Fractal Dimensions

A large number of methods have been developed for estimating fractal dimensions, such as box-count (Hall & Wood, 1993), variogram (Constantine & Hall, 1994; Kent & Wood, 1997), level-crossing (Feuerverger, Hall, & Wood, 1994), and spectral (Chan, Hall, & Poskitt, 1995) estimators. Essentially all methods follow a common scheme: (a) A certain numerical property Q of the data is computed as a function of scale ε; (b) an asymptotic power law $Q(\varepsilon) \propto \varepsilon^b$ is derived or postulated as the scale $\varepsilon \to 0$, where the scaling exponent b is a linear function of the fractal dimension D; and (c) applying linear regression of $logQ(\varepsilon)$ on $log\varepsilon$ returns an estimate of D.

Gneiting, Sevcikova, and Percival (2012) studied these estimation methods (see Appendix A for definitions) extensively in finite sample simulations. Considering both efficiency and robustness, the authors recommend the madogram estimator, a statistically more efficient version of the Hall – Wood estimator. Therefore, in the following analysis, only the madogram estimator of the fractal dimension is used.

Data and Fractal Analysis

In this study, a data set was obtained from the KDD Cup competition.[1] This data set is the Algebra I 2006 – 2007 training data that take the form of records of interactions between students and computer-aided tutoring systems. Students solve problems in the tutor, and each interaction between the student and computer is logged as a transaction. When using a computer tutor, a student completes a problem by steps. The whole collection of steps for one problem composes the solution, and the last step can be considered the answer. Students may not perform all steps for a problem. Students might complete a step by performing the correct steps, requesting a hint from the tutor, entering an incorrect value, or a combination of these. Each hint request, incorrect attempt, or correct attempt is a transaction, and each recorded transaction is referred to as an attempt for a step. The complete data record includes these attempts for a step, transaction time, step duration, correct first attempt (1 if correct, 0 otherwise), and so on.

In the data set studied here, there are 1,840 students and 2,270,384 steps in total. Hence, on average, each student took 1,234 steps throughout the online tutoring course.

The correct or incorrect first-attempt responses by students were tracked for each item or step before they requested hints. If a student requested a hint before providing any responses to a step, his or her response to the step was treated as incorrect. Let t denote the number of steps, $X(t) = 1$ if the first attempt is correct, and $X(t) = -1$ if incorrect. Let $Z(t)$ be the net score at time t, the accumulation of first attempts. That is,

$$Z(t) = \sum_{i=1}^{t} X(i).$$

When the $X(t)$ are independent and identically distributed, $Z(t)$ is a simple random walk. The graph of $\{(t, Z(t))\}$ is so irregular that its fractal dimension (both Hausdorff and box-counting dimensions) is 1.5 (Falconer, 1990, Theorem 16.4). When the $X(t)$ are strongly positively correlated or persistent (i.e., $1/2 < H \leq 1$), the fractal dimension of the graph is $2 - H$, a value smaller than 1.5. In this case, the graph is smoother. However, when the $X(t)$ are negatively correlated (antipersistent, $1 < H < 1/2$), the graph of $\{(t, Z(t)\}$ is very irregular so that it may have a fractal dimension approaching 2.

During students' learning processes, their step responses to online tutoring problems are expected to be autocorrelated because of knowledge- and content-based dependence. For instance, a good grasp of concepts in the previous sections may help in understanding concepts in the following sections. To determine how strong the autocorrelations are, and whether an autoregression moving average (ARMA) model or a self-similar process can be used to model the data, the dependence structure of these step response sequences is briefly investigated. For this purpose, plots of the sample autocorrelations against the lag and the variance of $\overline{X}_n$ against n on the log-log scale are considered (Beran, 1994).

Two students, Student 7 and Student 9, were picked from the pool for illustrative purposes. Based on the preliminary results of kernel regression, Figure 1 shows that Student 7's first attempts can be roughly assumed to be stationary, whereas Student 9's first attempts are stationary after the first 250 or so steps. Figure 2 shows the variance of $\overline{X}_n$ against n on the log-log scale of two students' net scores (for Student 9, variance estimation starts from the 251th step). The slopes (0.52 and 0.66) are less than 1, which indicate strong correlations in the observations (Beran, 1994). The ACF plots and the logarithm of absolute value of ACF against logarithm of lag are presented in Figure 3. The ACF plots on the first row show slow decay of autocorrelations for the two response sequences, and the logarithm of ACF against the logarithm of lag again shows the slow decay with a slope of −0.43 and −0.63 (a value between −1 and 0). Again, these ACF plots on the log-log scale indicate strong correlations in observations (Beran, 1994). Because of the evidence of strong correlation (long-range dependence), using an ARMA model with an excessive number of parameters is undesirable, especially because it increases the uncertainty of the statistical inference and hinders interpretation of the parameters. Therefore fractional dimensions were chosen to characterize dependency in the net score processes.

Figure 4 shows cumulative sums of correct (1) and incorrect (−1) scores in the sequences. For easy comparison, only the first 1,000 steps of the data were used. Notice that the scales are very different across the plots.

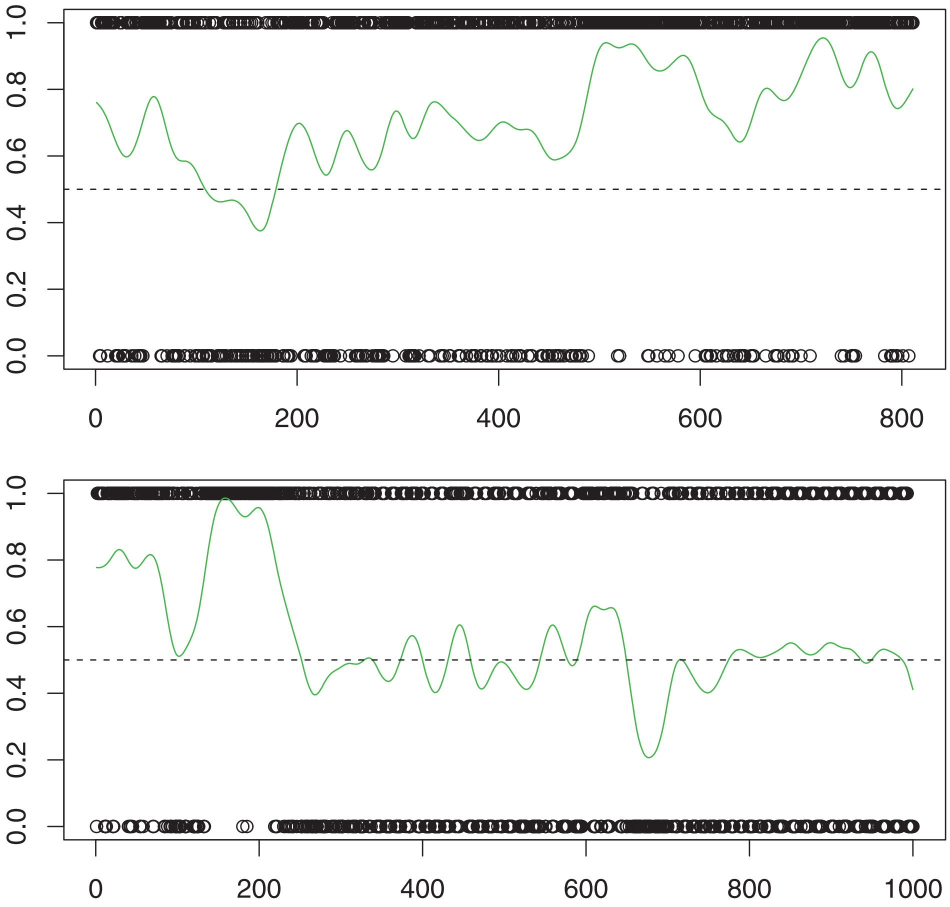


Figure 1 Score sequences of first attempts for Students 7 and 9. The curves in the middle are the kernel smoothing regression lines for response accuracy against number of steps. The *x* axis stands for the number of first attempts, and the *y* axis is the corresponding value (1 = correct response; 0 otherwise) of the first attempt.

To calculate the fractal dimension, the summands of the net score sequence were centered by removing the average observed score (a fixed number of each student) from the observed score to obtain a detrended sequence that had stationary increments. That is, a nonrandom linear trend was removed from the observed net scores. Theoretically, the net score sequences and detrended sequences have the same fractal dimension. In practice, they are not the same (but are somewhat close). By centering the responses, we are more focused on the dependence structure. The fractal dimensions of the detrended sequences are 1.37 for both Students 7 and 9.

In Figure 4, the trajectories of Student 7's and Student 9's cumulative responses have similar fractal dimensions around 1.37, but the trajectories are very different: Student 7's trajectory is consistently increasing, whereas Student 9's has dramatic changes. To analyze the behaviors in detail, Figure 5 shows the local fractal dimensions of the two net score sequences. The local fractal dimensions are calculated for every 100 observations (i.e., window width is 100) sliding through the whole sequence with a step width of 40.

In the upper panels of Figure 5, the solid line is a trajectory of the cumulative sum of responses (the net score path), the dashed line is the local fractal dimension, and a horizontal line of 1.5 is plotted at the center as a criterion line. The dash-dotted line is the overall (global) fractal dimension of the trajectory, and the two dotted lines are the estimated confidence band for the fractal dimension, estimated from the bootstrapping method (Davies & Hall, 1999). As discussed in Hall and Wood (1993), the fractal estimator approximates a normal distribution in the limit when $1.25 < D < 2$. The confidence band allows estimation error to be considered when evaluating the fractal dimension and persistency in students' responses. In the upper left panel of Figure 5, the fractal dimension fluctuates around 1.37, which may indicate a long-range dependence. That is, overall, the responses are positively correlated. When the fractal dimension is 1.5, it is the regular random walk. That is, the responses are independent of each other. For

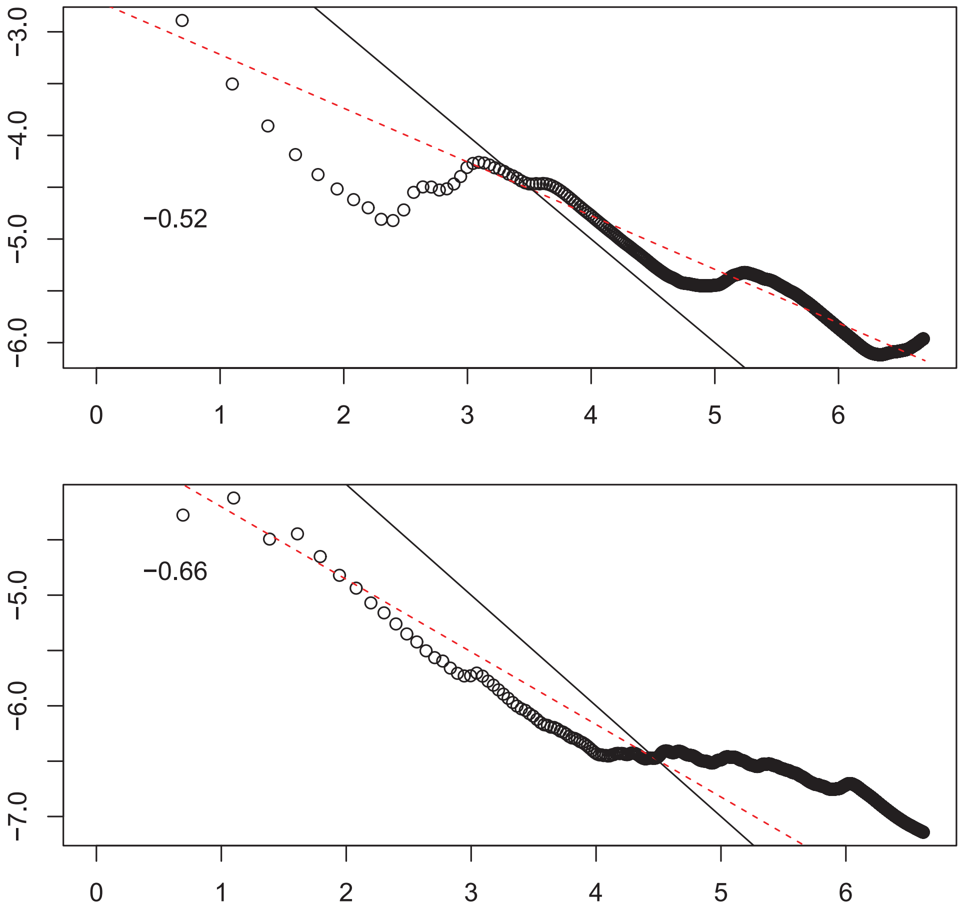


Figure 2 The log-log plot of variance of mean net scores of first attempts for Students 7 and 9. The *x* axis stands for *log logn*, and the *y* axis is the variance of $\overline{X}_n$, where *n* is the number of first attempts.

Student 9, as is shown in the upper right panel of Figure 5, the local fractal dimension reaches 1.2 around the 150th step or so, indicating a strong long-range dependence. However, it reaches 1.8 around the 300th step, which may indicate inconsistent performance. That is, from the 250th to 300th steps, the autocorrelation is negative. In some sense, the student's responses may be irregular during this period of learning, and a warning sign may be issued for early intervention. After 300 steps, the fractal dimension fluctuates around 1.4, indicating a persistent pattern in responses.

Fractal dimensions characterize dependency in the score processes, which is not necessarily linked to proficiency. However, to evaluate students' proficiency, knowledge, or skill in learning, the average/mean correct rate of responses may be a meaningful statistic to monitor. In the lower panels of Figure 5, the local mean net scores (i.e., the average net score of every 100 observations sliding through the whole sequence with a step width of 40) are plotted for the two net score sequences. In the plot, the solid line is the net score path, the dashed line is the local mean net score, and the dash-dotted line is the overall/global mean net score. The lower left panel shows that, for Student 7, the mean scores fluctuate about zero around the 100th to 200th steps and then rise to 0.5, showing a relatively good performance afterward. Combining with the upper left panel, the plots show that, overall, Student 7 persistently performed at a high proficiency level (with fractal dimensions less than 1.5 and mean net scores larger than zero). Conversely, the lower right panel in Figure 5 shows that Student 9 started with a high correct rate but dramatically dropped to the zero line (average net score of zero) around the 300th step. This change coincides with the high local fractal dimension in the upper right panel. Afterward, the local mean fluctuates around the zero line. Combining with the local fractal dimension, it shows that Student 9 had persistent performance but at a low proficiency level after the first 300 steps. A confidence band can be easily added to the mean score plots to account for uncertainty.

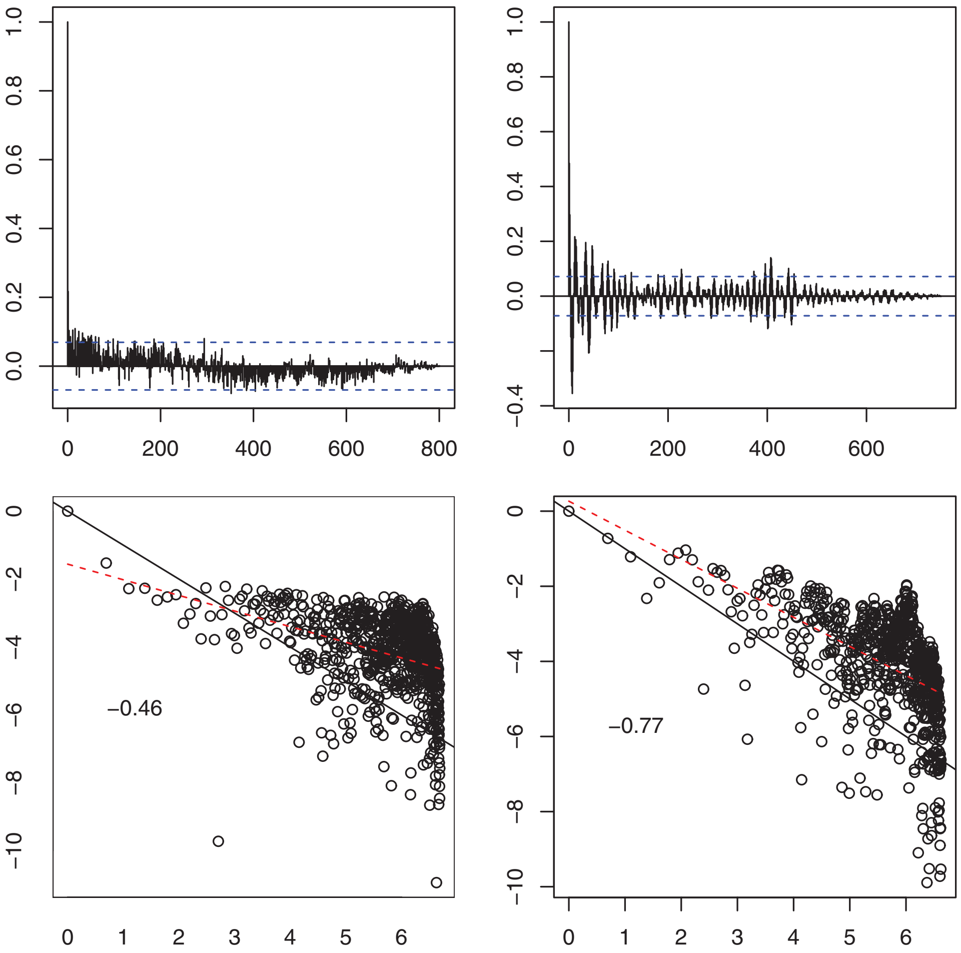


Figure 3 (top) Autocorrelation function and (bottom) logarithm of the autocorrelation function of the response accuracy sequences for Students 7 and 9. In the lower panels, the logarithm of the autocorrelation function (*y* axis) against the logarithm of lag (*x* axis) again shows the slow decay with a slope of −0.46 and −0.77 (dashed lines). The solid line is a line of slope −1.

Overall, both students showed persistence in their learning behavior with some instability at the first 200 or 300 steps, but one student had higher proficiency than the other. Figures like Figure 5, with the response trajectory, local mean, and local fractal dimension, may provide valuable information with regard to student learning behavior.

Discussion

As pointed out by Winne and Baker (2013), in educational data mining, instruments should be developed to gather data that trace over time to inform how students are learning. With the popularity of online learning/tutoring and the abundance of online processing data, there is a need to understand basic structures of online learning data.

In this study, data collected from an online tutoring system for individual students were explored. Strong dependence was observed in some of the students' first-attempt responses to questions. It was shown that the dependence of online data might be characterized by fractal dimensions as a summary statistic locally and globally. For students' net score path, a fractal dimension close to 1 indicates a strong positive correlation among their responses and therefore a persistent behavior in the learning process; a fractal dimension close to 1.5 indicates no correlation among responses and that their behavior may be subjected to random guessing; and a fractal dimension close to 2 indicates irregular and antipersistent behavior in responses, which may be a warning sign that the student is struggling in the learning process. In short, the fractal dimension may reflect persistence in students' response behavior.

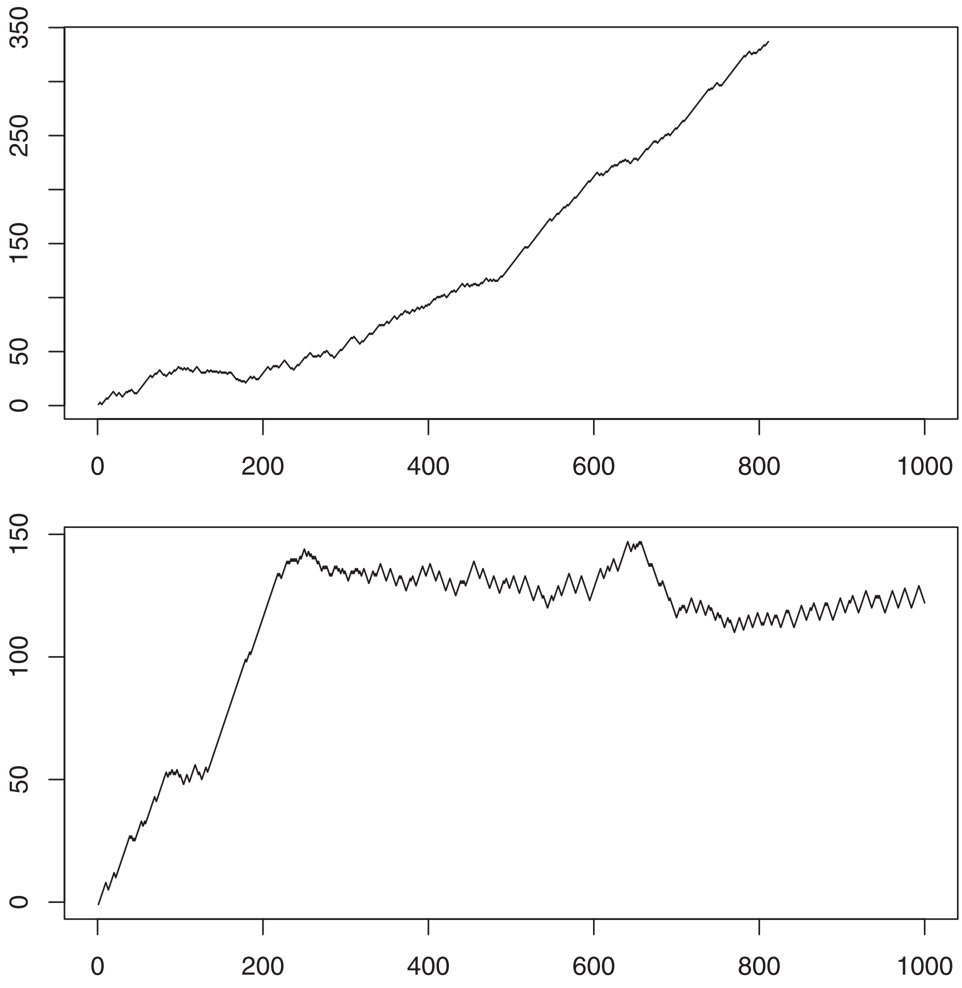


Figure 4 Net score paths for the two students. The *x* axis stands for the number of first attempts, and the *y* axis is the cumulative sum score.

It may be useful for researchers and instructors to monitor the global and local mean scores, the global and local fractal dimensions, and their confidence bands of an individual student's responses. The combined information, particularly sudden changes, may reflect changes in an individual student's learning behavior in short and long learning windows, and it may indicate whether the student is struggling and needs extra help in the learning process.

However, the results obtained from fractal analysis are experimental and inconclusive. Many questions are unanswered, for example, how to set a threshold to flag the local fractal dimension, that is, when it is high enough to warrant a warning or flagging. This issue may be addressed with external variables, such as assessment scores. Choice of window size in calculating local fractal dimensions and local mean scores should depend on numbers of steps and questions in each knowledge content. Another caution is that the time-series techniques (including estimation of the fractal dimensions) are best suited to long and stationary sequences. Short sequences will unavoidably cause problems in estimation accuracy, particularly when trends are presented. Further studies also need to consider whether a multidimensional stochastic process can model response accuracy and response time simultaneously so that more useful information can be extracted from the online data. In addition, the fractal analysis of online data presented here mainly addresses persistence in students' response behavior. To reflect students' overall proficiency, global and local mean score analysis may be helpful, and the confidence band is likely to be wider if a strong dependence is observed. However, for researchers who are interested in students' particular skills and content knowledge, more educational, psychological, and statistical models need to be employed.

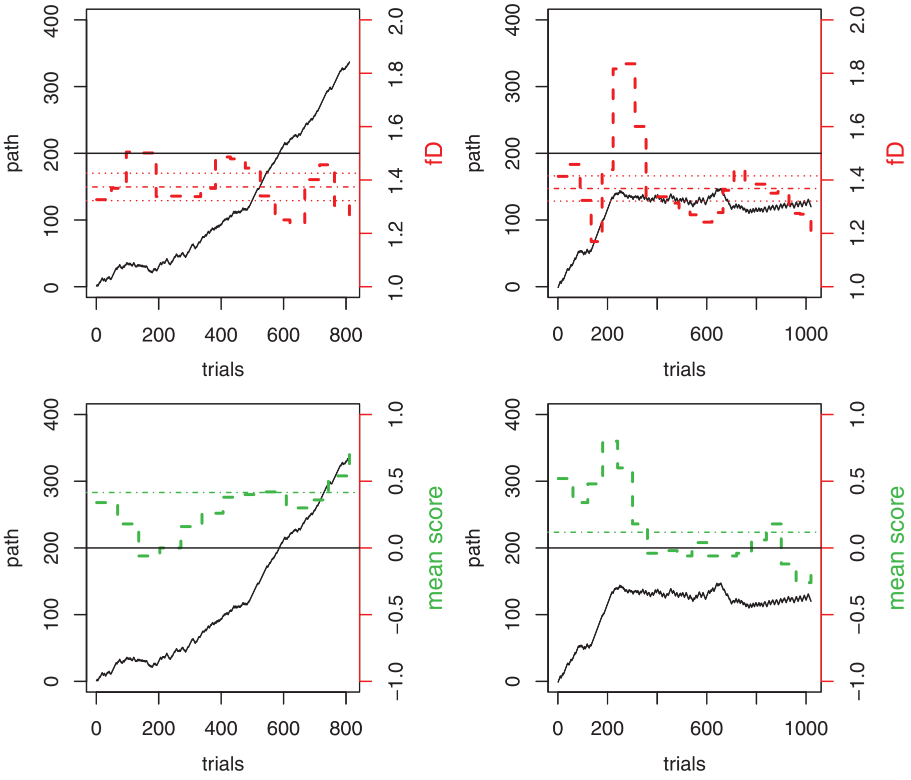


Figure 5 Local dimensions (first row) and local mean net scores (second row) of Students 7 and 9 (window size = 100, step size = 40). A low dimension close to 1 and less than 1.5 indicates long-range dependence; 1.5, random walk; and 2, antipersistency. Local mean net scores close to zero indicate an equal chance of correct and incorrect answers.

Acknowledgments

The author is very grateful to Rebecca Zwick, Yoav Bergner, and other reviewers. She is also thankful to Kim Fryer for editing the report.

Note

[1] See https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp/

References

Beran, J. (1994). *Monographs on statistics and applied probability: Vol. 61. Statistics for ling-memory processes.* New York, NY: Chapman and Hall.

Chan, G., Hall, P., & Poskitt, D. S. (1995). Periodogram-based estimators of fractal properties. *Annals of Statistics, 23*, 1684–1711.

Constantine, A. G., & Hall, P. (1994). Characterizing surface smoothness via estimation of effective fractal dimension. *Journal of the Royal Statistical Society: Series B, 56*, 97–113.

Davies, S., & Hall, P. (1999). Fractal analysis of surface roughness by using spatial data. *Journal of the Royal Statistical Society: Series B, 61*, 3–37.

Falconer, K. (1985). *The geometry of fractal sets*. Cambridge, England: Cambridge University Press.

Falconer, K. (1990). *Fractal geometry*. Chichester, England: John Wiley.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine, 17*, 37–54.

Feuerverger, A., Hall, P., & Wood, A. T. A. (1994). Estimation of fractal index and fractal dimension of a Gaussian process by counting the number of level crossings. *Journal of Time Series Analysis, 15*, 587–606.

Gneiting, T., Sevcikova, H., & Percival, D. (2012). Estimators of fractal dimensions: Assessing the roughness of time series and spatial data. *Statistical Science, 27*, 247–277.

Hall, P., & Wood, A. (1993). On the performance of boxcounting estimators of fractal dimension. *Biometrika, 80*, 246–252.

Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). *HarvardX and MITx: The first year of open online courses* (HarvardX and MITx Working Paper No. 1). https://doi.org/10.2139/ssrn.2381263

Kent, J. T., & Wood, A. T. A. (1997). Estimating the fractal dimension of a locally self-similar Gaussian process by using increments. *Journal of the Royal Statistical Society: Series B, 59*, 679–699.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mandelbrot, B. (1983). *The fractal geometry of nature*. New York, NY: Macmillan.

Taqqu, M. (2003). Fractional Brownian motion and long-range dependence. In P. Doukhan, G. Oppenheim, & M. Taqqu (Eds.), *Long-range dependence* (pp. 5–38). Boston, MA: Birkhauser.

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometricka, 72*, 287–308.

Warnakulasooriya, R., & Galen, W. (2012, June). *Categorizing students' response patterns using the concept of fractal dimension*. Paper presented at the Educational Data Mining conference, Chania, Greece. Retrieved from http://educationaldatamining.org/EDM2012/uploads/procs/Posters/edm2012poster17.pdf

Winne, P., & Baker, R. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining, 5*, 1–8.

Appendix A

Computing the Fractal Dimension

The mathematical definition of fractal dimension is based on the measure theory (counting measure, Lebesgue measure as an extension of length, probability as measure, etc.). Hausdorff dimension dim_H is based on the Hausdorff measure; packing dimension dim_P is based on the packing measure (Falconer, 1985, 1990).

The Hausdorff dimension is an extension of the traditional dimension, as the Hausdorff measure is an extension of the traditional Lebesgue measure, which in turn is an extension of length. Many methods exist to compute the fractal dimension; each method has its own theoretic basis. This fact often leads to obtaining different dimensions by different methods for the same set.

Let $F \subset \mathcal{R}^d$ be a set. For $\varepsilon > 0$, an ε-cover of F is a finite or countable collection $\{B_i : i = 1,2,\cdots\}$ of balls $B_i \subset \mathcal{R}^d$ of diameter $\|B_i\|$ less than or equal to ε that covers F. Let

$$H^{\delta}(F) = lim \inf_{\epsilon \to 0} \left\{ \sum_{i=1}^{\infty} \left\|B_i\right\|^{\delta} \; : \; \left\{B_i \; : \; i = 1,2,\cdots\right\} \text{ is an } \epsilon\text{-cover of } F \right\} \tag{A1}$$

denote the δ-dimensional Hausdorff measure of F. Because of monotonicity of the Hausdorff measure with respect to δ, there exists a unique nonnegative value D such that $H^{\delta}(F) = \infty$ if $\delta < D$ and $H^{\delta}(F) = 0$ if $\delta > D$. This value D is called the Hausdorff dimension $\dim_H = \mathrm{D}$ of the point set F.

The box-counting dimension is widely used and easily computed (but not based on the measure theory). Let F be any nonempty bounded subset of $\mathcal{R}^n$, and let $N_\delta(F)$ be the smallest number of sets of diameter at most δ that can cover F. The lower and upper box-counting dimensions of F, respectively, are defined as

$$\underline{\dim}_B F = \underline{lim}_{\delta \to 0} \frac{logN_{\delta}(F)}{-log\delta}$$

$$\overline{\dim}_B F = \overline{lim}_{\delta \to 0} \frac{logN_{\delta}(F)}{-log\delta}.$$

If the above two limits are equal, the common value is the box-counting dimension of F.

Note that $\dim_H F \leq \underline{\dim}_B F \leq \overline{\dim}_B F$ and $\dim_H F \leq \overline{\dim}_B F$. For many regular fractals, the dimensions are the same.

Of course, there are mathematic and probabilistic techniques to compute the fractal dimension. The simplest example is the Cantor set (Falconer, 1985, p. 58). Let $m \geq 2$ be an integer and $0 < \lambda < 1/m$. Let F be the set obtained by the construction in which each basic interval I is replaced by m equally spaced subintervals of length $\lambda|I|$, the ends of I coinciding with the ends of the extreme subintervals. Then $\dim_H F = \dim_B F = \frac{logm}{-log\lambda}$. For example, the Cantor set in Figure A1 is $log2/log3$.

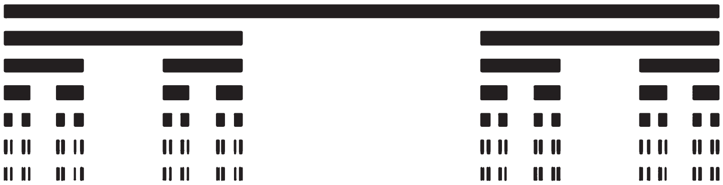


Figure A1 The Cantor set.

In this section, the estimators of fractal dimension in the R package *fractaldim* are reviewed. In their paper, Gneiting et al. (2012) restricted their attention to the point set

$$Z = \left\{ \left(t, Z_t\right) \in \mathcal{R}^d \times \mathcal{R} \, : \, t \in T \subset \mathcal{R}^d \right\} \subset \mathcal{R}^{d+1},$$

which is the graph of time series or spatial data observed at a finite set $T \subset \mathcal{R}^d$. Without loss of generality, T is assumed to be the unit interval or unit cube. For a smooth and differentiable curve ($d = 1$) or surface ($d = 2$), its fractal dimension, D, equals the topological dimension d. For a rough and nondifferentiable curve or surface, the fractal dimension may exceed the topological dimension. For example, for a Gaussian process $\{Z_t, t \in \mathcal{R}^d\}$ with stationary increments, if the covariance structure (variogram)

$$\gamma_2 (t) = \frac{1}{2} E \left(Z_\mu - Z_{\mu+t}\right)^2 \tag{A2}$$

satisfies

$$\gamma_2 (t) = \left\| c_2 t \right\|^\alpha + O\left(\|t\|^{\alpha+\beta}\right), \quad \text{as} \quad t \to 0, \tag{A3}$$

where $\alpha \in (0,2]$, $\beta > 0$, and $c_2 > 0$, then the graph of a sample path has a fractal dimension of

$$D = d + 1 - \frac{\alpha}{2} \tag{A4}$$

almost surely (Gneiting et al., 2012). Comparing Equation 1 to Equation A4, one can see that $\alpha = 2H$.

Table A1 lists the most popular estimators of the fractal dimension.

Box-Counting Estimator

As mentioned earlier, the box-counting estimator is the simplest and most popular method. In this R package, the basic idea is simple: First, the time series graph is initially covered by a single box; second, the box is divided into four quadrants, and the number of cells required to cover the curve is counted; third, each subsequent quadrant is divided into four subquadrants; and fourth, one continues doing so until the box width equals the resolution of the data, keeping track of the number of quadrants required to cover the graph at each step. The box-counting estimator equals the slope in an ordinary least squares regression of $logN(\varepsilon)$ on $log\varepsilon$, where ε is the width of the box and $N(\varepsilon)$ is the number of boxes covering the curve.

In the R package, the smallest scales k for which $N(\varepsilon_k) > n/5$ (where $\varepsilon_k = 2^{k-K}$ for $k = 0,1, \cdots, K$, and n is the sample size), as well as the two largest scales, are excluded from the regression fit.

Hall–Wood Estimator

The next three estimators (Hall–Wood, variogram, and variation) use the local property in Equation 4 of the time series in the time domain.

Table A1 Some Methods for Estimating the Fractal Dimension

Method	Property	Scale	Scaling law	Regime
Box count	$N(\varepsilon)$: number of boxes	ε: box width	$N(\varepsilon) \propto \varepsilon^{-D}$	$\varepsilon \to 0$
Variogram	$\gamma_2(t)$: variogram	t: lag	$\gamma_2(t) \propto t^{4-2D}$	$t \to 0$
Madogram	$\gamma_1(t)$: madogram	t: lag	$\gamma_1(t) \propto t^{2-D}$	$t \to 0$
Spectral	$f(\omega)$: spectral density	ω: frequency	$f(\omega) \propto \omega^{2D-5}$	$\omega \to 0$
Wavelet	$v^2(\tau)$: wavelet variance	τ: scale	$v^2(\tau) \propto \tau^{4-2D}$	$\tau \to 0$

Let $A(\varepsilon)$ denote the total area of the boxes at scale ε that intersect with the data graph. There are $N(\varepsilon)$ such boxes, and so $A(\varepsilon) = \varepsilon^2 N(\varepsilon)$. Therefore

$$\dim_B = 2 - lim \frac{logA(\varepsilon)}{log\varepsilon}. \tag{A5}$$

At scale $\varepsilon_l = l/n$ for $l = 1,2, \cdots, n$, an estimator of $A(l/n)$ is

$$\widehat{A}(l/n) = \frac{l}{n}\sum_{i=1}^{[l/n]} |Z_{il/n} - Z_{(i-1)l/n}|.$$

The Hall–Wood estimator only uses two points at the smallest scales:

$$\widehat{D}_{hw} = 2 - \frac{log\widehat{A}(2/n) - log\widehat{A}(1/n)}{log2}. \tag{A6}$$

Variogram Estimator

The variogram method is easy to implement. In view of Equations A2 and A3, the classical method of moments estimator is used for $\gamma(t)$ at lag $t = l/n$, defined as

$$\widehat{V}_2(l/n) = \frac{1}{2(n-l)}\sum_{i=1}^{n} \left(Z_{i/n} - Z_{(i-1)/n}\right)^2. \tag{A7}$$

The variogram estimator is the slope of the regression fit of $log\widehat{V}(t)$ on $logt$. As in the Hall–Wood estimator, to avoid bias, the implemented variogram estimator is

$$\widehat{D}_{V;2} = 2 - \frac{1}{2}\frac{log\widehat{V}_2(2/n) - log\widehat{V}_2(1/n)}{log2}. \tag{A8}$$

Variation Estimator

It is a generalized variogram estimator so that the estimator is more robust compared to the moments estimators of variogram. Define the variogram of order p of a stochastic process with stationary increments,

$$\gamma_p(t) = \frac{1}{2}E\left\|Z_u - Z_{t+u}\right\|^p. \tag{A9}$$

When $p = 1$, it is called madogram. It is shown that, for some constant c_p,

$$\gamma_p(t) = |c_p t|^{\alpha p/2} + O\left(|t|^{(\alpha+\beta)p/2}\right), \quad as \quad t \to 0.$$

The implemented variation estimator is

$$\widehat{D}_{V;p} = 2 - \frac{1}{p}\frac{log\widehat{V}_p(2/n) - log\widehat{V}_p(1/n)}{log2} \tag{A10}$$

for appropriately defined $\widehat{V}_p$.

It is observed in literature that $p = 1$ (the madogram estimator) is optimal most of the time in simulations.

Spectral and Wavelet Estimators

Both spectral (Whittle estimator) and wavelet estimators use the global property of Equation 4 of the time series, but in the frequency domain. For a stationary Gaussian process $\{X_t : t \in [0, 1]\}$, define the semiperiodogram

$$J(\omega) = B(\omega)^2, \qquad B(\omega) = 2\int_0^1 X_t cos(\omega[2t-1])\,dt.$$

Suppose there are $n = 2m + 1$ observations at time $t = i/(2m) \in [0, 1]$. The semiperiodogram estimator is

$$\hat{D}_p = \frac{5}{2} + \frac{1}{2}\left\{\sum_{l=1}^{L}(s_l - \bar{s}log(\hat{J}\left(\omega_l\right)\right\}\left\{\sum_{l=1}^{L}\left(s_l - \bar{s}\right)^2\right\}^{-1}, \tag{A11}$$

where $\omega_l = 2\pi l$, $s_l = log\omega_l$ and L is recommended to be $min\{m/2, n^{2/3}\}$.

Considering both efficiency and robustness, Gneiting et al. (2012) recommended the use of the madogram estimator, which can be interpreted as a statistically more efficient version of the Hall–Wood estimator.

Appendix B

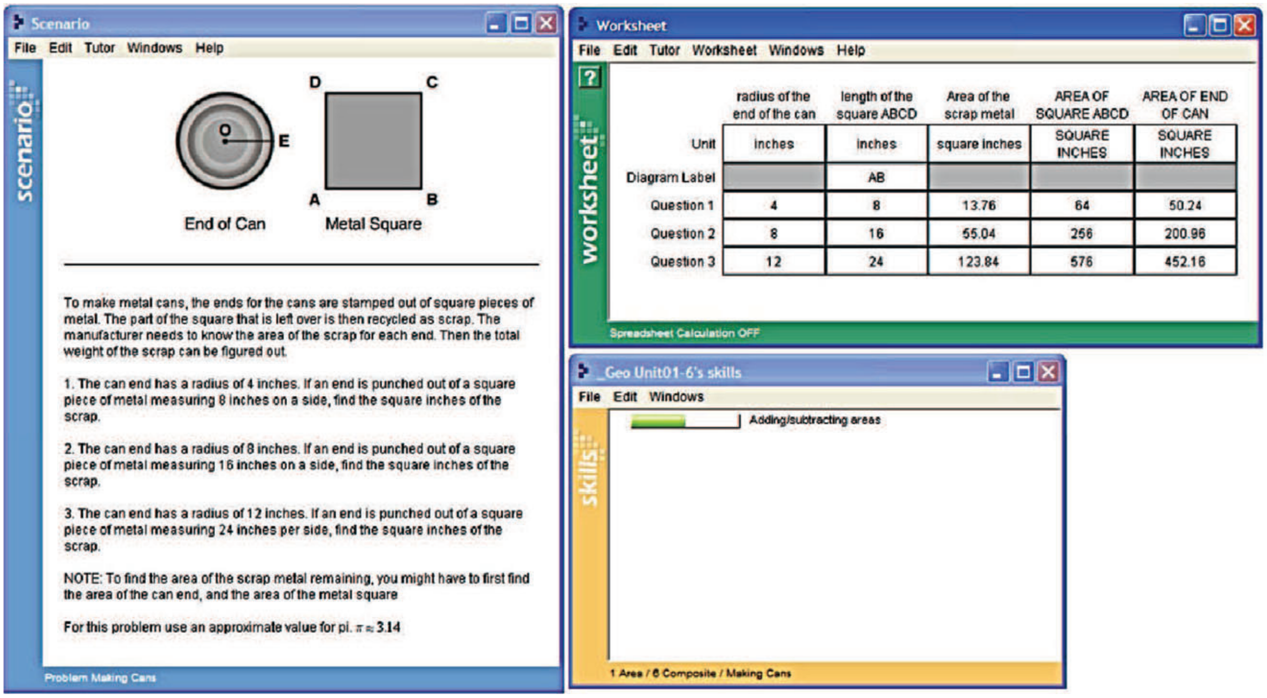
An Example Question in the Studied Data

The data take the form of records of interactions between students and computer-aided tutoring systems. The students solve problems in the tutor. Figure B1 is an exemplary scenario of the online questions.

In the example, a student is asked to find the area of a piece of scrap metal left over after removing a circular area (the end of a can) from a metal square (Figure B1). The student enters everything in the worksheet, except for the row labels and the column and unit labels for the first three columns. There are three questions. A problem is a task for a student to perform that typically involves multiple steps. A step is an observable part of the solution to a problem. For example, for the first question, there are five steps in the interface:

1. Find the radius of the end of the can (a circle).
2. Find the length of the square ABCD.
3. Find the area of the end of the can.
4. Find the area of the square ABCD.
5. Find the area of the leftover scrap.

This whole collection of steps composes the solution. The last step can be considered the answer, and the others are intermediate steps.



To make metal cans, the ends for the cans are stamped out of square pieces of metal. The part of the square that is left over is then recycled as scrap. The manufacturer needs to know the area of the scrap for each end. Then the total weight of the scrap can be figured out.

1. The can end has a radius of 4 inches. If an end is punched out of a square piece of metal measuring 8 inches on a side, find the square inches of the scrap.

2. The can end has a radius of 8 inches. If an end is punched out of a square piece of metal measuring 16 inches on a side, find the square inches of the scrap.

3. The can end has a radius of 12 inches. If an end is punched out of a square piece of metal measuring 24 inches per side, find the square inches of the scrap.

NOTE: To find the area of the scrap metal remaining, you might have to first find the area of the can end, and the area of the metal square

For this problem use an approximate value for pi. $\pi \approx 3.14$

	radius of the end of the can	length of the square ABCD	Area of the scrap metal	AREA OF SQUARE ABCD	AREA OF END OF CAN
Unit	inches	inches	square inches	SQUARE INCHES	SQUARE INCHES
Diagram Label		AB			
Question 1	4	8	13.76	64	50.24
Question 2	8	16	55.04	256	200.96
Question 3	12	24	123.84	576	452.16

Figure B1 A problem from Carnegie Learning's Cognitive Tutor Geometry.

Note that students might not (and often do not) complete a problem by performing only the correct steps. Instead, the student might request a hint from the tutor or enter an incorrect value.

There are many variables in the data set, such as problem, step, knowledge component, and opportunity. As a preliminary analysis, this report only focused on one variable: correct first attempt, the tutor's evaluation of the student's first attempt on the step—it is 1 if correct, 0 if an error.

Suggested citation:

Guo, H. (2017). *Exploring online learning data using fractal dimensions* (Research Report No. RR-17-15). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12143

Action Editor: Rebecca Zwick

Reviewers: Yoav Bergner and Kentaro Yamamoto