980624

Symbolic Time-Series Analysis of Engine Combustion Measurements

C.E.A. Finney
University of Tennessee

J.B. Green, Jr., **C.S. Daw**
Oak Ridge National Laboratory

ABSTRACT

We present techniques of symbolic time-series analysis which are useful for analyzing temporal patterns in dynamic measurements of engine combustion variables. We focus primarily on techniques that characterize predictability and the occurrence of repeating temporal patterns. These methods can be applied to standard, cycle-resolved engine combustion measurements, such as IMEP and heat release. The techniques are especially useful in cases with high levels of measurement and/or dynamic noise. We illustrate their application to experimental data from a production V8 engine and a laboratory single-cylinder engine.

MOTIVATION

Recent studies have demonstrated that cyclic combustion variations in spark-ignition engines under lean fueling exhibit patterns that can be explained as the result of noisy nonlinear combustion instabilities [1, 2, 3, 4]. These instabilities are dominated by the effects of residual cylinder gas (prior-cycle effects) and noisy perturbations of engine parameters. Because this dynamical noise obscures the underlying deterministic patterns, it is difficult to observe changes in these patterns as nominal engine parameter values are changed. This difficulty is often increased by inaccuracies in measurement of characteristic combustion parameters such as IMEP and heat release.

Symbolization of engine combustion measurements provides a simple, yet effective way to minimize the effects of dynamic noise and measurement error so that deterministic prior-cycle effects can be quantified more accurately. If applied at the time of data acquisition, symbolization also becomes a means of data compression and greatly speeds data processing. These characteristics make data symbolization a useful tool for on-board engine diagnostics and real-time control. Because of this potential value, we have begun studying how the symbol-statistics approach to time-series analysis can be applied in the context of practical automotive engines.

EXPERIMENTAL APPARATUS

The first experimental engine was a production Ford 4.6-L V8. The nominal engine operating condition was 1200 RPM, 27.1 $\mathrm{N \cdot m}$ brake torque, and 20 CAD before top center spark. A dynamometer was operated in speed-control mode to maintain nearly constant engine speed despite erratic combustion at very lean conditions. Feedback engine controllers were engaged to achieve an operating condition; once the condition was achieved, the feedback controllers were shut off, and the engine was run in open-loop mode, except for dynamometer speed control. This strategy assured that combustion was minimally influenced by feedback controllers while the engine ran at constant speed.

Combustion pressure was recorded once per CAD from a single cylinder and nominal operating conditions at a 50-Hz rate for over 2800 contiguous cycles. Combustion pressure measurements were made with a piezo-electric pressure transducer mounted in the cylinder head. To provide a dynamic measurement that could be compared with the model, the combustion heat release for each cycle was calculated by integrating the cylinder pressure data using a technique equivalent to the Rassweiler-Withrow method [5, 6]. As a result, for each engine experiment a time series of over 2800 heat-release values was produced.

The second experimental engine system was a single-cylinder spark-ignited, L-head engine (Kohler Magnum 12). The engine has a 0.476-L displacement, 8.59-cm bore, 8.26-cm stroke and compression ratio of 6.6. The fuel used in all experiments was indolene. Three thousand cycles of in-cylinder pressure data were taken on a crank-angle-resolved basis for the experiments. For these tests, equivalence ratio was varied from 0.99 to 0.72. This was achieved by keeping the throttle position constant at 11 percent of the value at wide-open throttle and adjusting the injected fuel pulse width. Spark timing was maintained at forty degrees before top center based on the maximum brake torque at that throttle setting and stoichiometric fueling conditions. A motoring dynamometer was used to maintain an engine speed of 1200 RPM, even during very lean

operating conditions. All other engine controls were disabled to minimize the effect of these controls on cycle-by-cycle engine dynamics.

DEFINITIONS OF SYMBOLIC TIME-SERIES ANALYSIS

We employ the symbol-statistics approach to time-series analysis suggested by Tang *et al.* [7]; an alternate but analogous approach may be found in [8]. Although this method is motivated by symbolic-dynamics theory, it is not completely rigorous, mainly because generating partitions are undefined in the presence of noise (a generating partion is a division of the data space which yields a perfect symbolic description of the time-series structure; see [9] for a detailed discussion). We adopt and define new terminology and nomenclature to avoid confusion among readers familiar with formal symbolic-dynamics theory, although we provide the corresponding terms as parenthetical statements wherever appropriate. In doing so, we hope to remind readers that although this methodology is derived from symbolic dynamics, in place of theory it must rely on empirically and heuristically motivated arguments.

SYMBOLIZATION AND SYMBOL-SEQUENCE HISTOGRAMS — Data symbolization involves the conversion of a data series of many possible values into a symbol series of only a few distinct values. This "coarse-graining" has the practical effect of producing low-resolution data from high-resolution data; only large-scale features are captured, and the effects of dynamic and measurement noise on statistical algorithms are reduced. In the practical sense for studying engine dynamics, only a qualitative description, such as "stronger" combustion or "weaker" combustion, might be desired, with many of the intermediate levels neglected.

A collection of data sampled over time from some physical measurement forms a time series. Time series may represent a measurement continuously over time, or they may represent measurements from discrete events. For instance, cylinder pressure measured multiple times throughout an engine cycle through consecutive cycles would represent a continuous time series, whereas the maximum cylinder pressure in each engine cycle over consecutive cycles would represent a discrete time series. For the present work, we discuss only discrete time series.

The basic idea in data symbolization is to discretize time-series values into a few possible values. Depending on the value of a given datum, it is assigned one of n symbolic values (*e.g.*, 0 or 1 for $n = 2$). We refer to n as the *symbol-set size*, which denotes the number of symbols available to symbolize the data; in the symbolic-dynamics literature, n is referred to as the *alphabet size*. As n becomes increasingly large, more detail of the original time series is included, with the tradeoff that more noise is included. In the limit, when n equals the number of distinct values in the time series, the symbol series and the time series contain the same degree of precision, with the data representation (*e.g.*, using symbols instead of numbers) being the only difference.

For any symbol-set size, the choice of how to partition the data space affects the characteristics of the symbolic description of the data. Although some researchers attempt to optimize partitioning [8], it is typical to define symbolization partitions such that the individual occurrence of each symbol is equiprobable with all others [7]. Thus, for a symbol-set size of 2, the data median defines the partition between symbols.

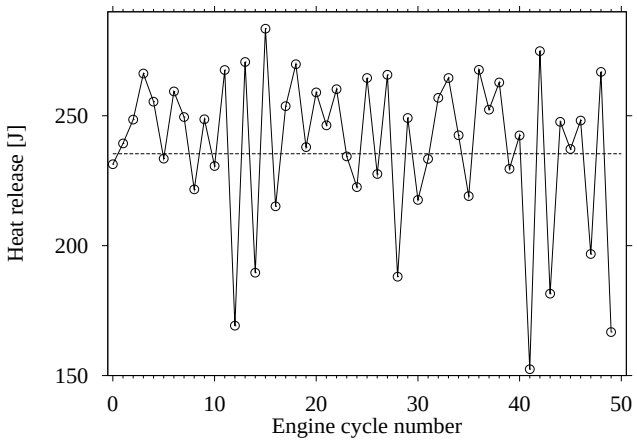

Figure 1: Example time series with symbol partition.

Figure 1 shows a combustion heat-release data series from the V8 engine at a nominal equivalence ratio of 0.59. Visibly, the heat release varies erratically over consecutive cycles. The dashed line represents the symbol partition — data below the partition are symbolized as 0, and data above the partition are symbolized as 1. As an illustration, in this figure, for groupings of six consecutive engine cycles, data points 0–5 form the symbol sequence 011110, points 1–6 form the symbol sequence 111101, and so on.

Once a symbol series is generated, the *symbol statistics*, reflecting the probability of occurrence for different symbol sequences, may be calculated. The *symbol tree* is a graphical representation of the symbol statistics as a function of the symbol-sequence length. An example tree is shown below:

$$\begin{array}{cccccccc} & & p_0 & & & & p_1 & \\ & p_{00} & & p_{01} & & p_{10} & & p_{11} \\ p_{000} & & p_{001} & & p_{010} & & p_{011} & \\ & p_{100} & & p_{101} & & p_{110} & & p_{111} \end{array}$$

In the bottom level of the tree, p_{xyz} is the probability that the symbol sequence xyz occurs. In this example, the symbol-sequence length, denoted L, is 3, representing the probability of 3 successive symbols to occur. Some researchers refer to L as the tree level because it represents the depth in the symbol tree at which the symbol statistics are evaluated. Typically, the symbol statistics vary as a function of the symbol-sequence length.

A convenient numerical representation of each symbol sequence is achieved by converting the base-n sequence into a base-10 equivalent number, which we term the *sequence code*. For instance, for the base-2 sequence 000 the sequence code is 0, for 001 is 1, for 010 is 2, for 011 is 3 and so on. This is very similar to the approach used by Lehrman, Rechester and White [8]. Using this method of identifying each symbol sequence also allows for the symbol statistics to be displayed as a histogram, with the sequence code serving as the bin number; we refer to this representation of the symbol statistics as a *symbol-sequence histogram*. Because of our equiprobable partitioning rule, the relative frequencies for truly random data sequences will be equal, and all histogram bins will be equally probable (subject to the availability of sufficient data). Thus,

any significant deviation from equiprobability is indicative of time correlation and deterministic structure.

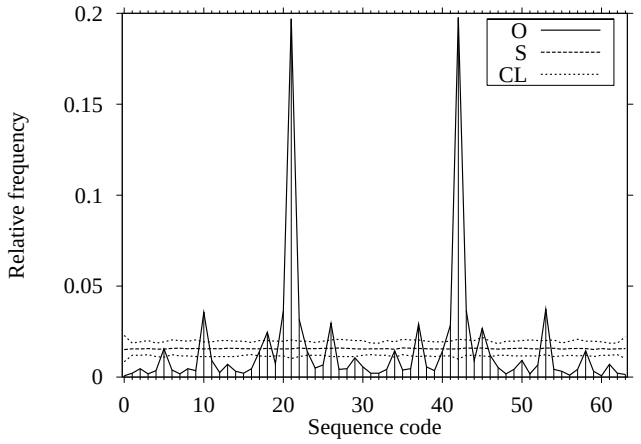

Figure 2: Example symbol-sequence histogram. Solid lines ("O") represents the original symbol series, and dashed lines represent the mean ("S") and 95 percent confidence limits ("CL") for 200 shuffled surrogate series.

Figure 2 shows the symbol-sequence histogram for 2843 engine cycles from the complete time series from which the segment in **Fig. 1** was extracted (the significance of shuffled surrogates is discussed below). The dominant symbol patterns are for sequence codes 21, representing the symbol sequence 010101, and 42, representing the symbol sequence 101010. Thus, the most prevalent feature of the symbol series is an oscillation between 0 and 1; physically, this corresponds to alternating larger and smaller heat releases.

Note that in many cases, the symbol-sequence histogram peaks are related. For instance, symbol sequence 101010 or 101011 must follow sequence 010101 in time. As a result, when 010101 is the dominant pattern, as seen in **Fig. 2**, the relative frequencies of sequences 101010 and 101011 are high. At higher symbolization levels (*e.g.*, $n > 2$), frequency mismatches between sequences which are time inverses of each other (*e.g.*, 0312 and 2130 for $n = 4$) reveal time irreversibility. Such time-irreversibility features reveal an "arrow of time" in the data, a key characteristic of nonlinear processes [10]. Even with the binary symbolization seen in **Fig. 2**, the symbol-sequence histogram is asymmetric about its axis (*e.g.*, note that the frequencies of sequence code 21 and 42 differ as well as those for sequence codes 10 and 53), indicating that certain sequences are favored over their complements.

OPTIMIZATION OF SYMBOLIZATION PARAMETERS — The average degree of organization of a symbol-sequence histogram at any tree level is implicit in the Shannon entropy. Similar to Tang *et al.* [7], we define a modified Shannon entropy as:

$$H_S(L) \equiv -\frac{1}{\log N_{\rm seq}} \sum_i p_{i,L} \log p_{i,L} \qquad (1)$$

where $N_{\rm seq}$ is the total number of sequences with non-zero frequency, i is a string-sequence index, and $p_{i,L}$ is the probability of string sequence i of length L. The only difference between **Eq. 1** and the definition used by Tang *et al.* is that we use the number of non-zero frequency sequences rather than the total number of possible sequences. For random data H_S should equal 1, whereas for nonrandom data it should be between 0 and 1. In the current context, lower H_S implies more deterministic structure.

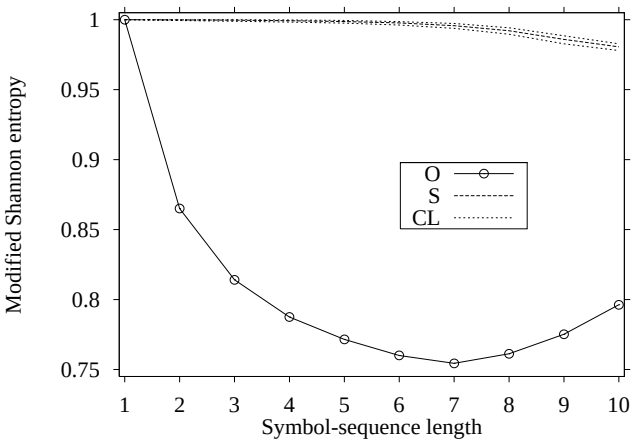

Figure 3: Modified Shannon entropy vs. symbol-sequence length for noisy engine heat-release data. Solid lines ("O") represents the original symbol series, and dashed lines represent the mean ("S") and 95 percent confidence limits ("CL") for 200 shuffled surrogate series.

The value of H_S varies depending on the symbol-sequence length L. There is no theoretical rule to determine which L is sufficient or best for observing significant sequence patterns, if they exist. However, we have developed an empirical approach to select an appropriate L using the modified Shannon entropy. Specifically, we find that H_S typically reaches a minimum value as L is increased from 1. This trend is illustrated in **Fig. 3** using the V8 experimental data used in **Fig. 1** and **2** (the significance of shuffled surrogates is discussed below). These data contain deterministic combustion patterns obscured by significant noise. As the deterministic component is better defined by a progressively longer L, the nonrandom part of the engine model becomes more evident, even though significant parameter noise is present. We explain this minimum in H_S as reflecting the symbol-sequence transformation which best distinguishes the data from a random sequence. Symbol sequences that are too short lose some of the important deterministic information; symbol sequences that are too long reflect noise and data depletion (*i.e.*, there are not enough data to get reliable statistics for such long sequences). Thus one can argue that the L value for which H_S is minimum is an "optimal" choice for the given data and symbol-set size. For convenience, in most of the examples below, we choose to use a symbol-sequence length of 6, which is optimal or nearly optimal for a wide variety of experimental conditions and also yields a moderate number of possible sequences (64 for $n^L = 2^6$).

The significance of the symbol statistics depends on time-series length in relation to the symbolization parameters. When the number of data is on the order of the number of possible sequence codes (related as n^L), data are too sparse for accurate statistical characterization of the data, particularly in distinguishing with confidence nonrandom from random features in the data. For moderate-to-large data-set sizes, it is possible

to establish confidence limits on the significance of the symbol statistics and their subsidiary statistics (*e.g.*, H_S). One method is to compare symbol statistics from the original time series with statistics from shuffled surrogates. Here, shuffled surrogate means that the order of the original time series is shuffled randomly. If the original time series have significant sequence patterns, then the shuffling process should destroy those patterns, and the symbol statistics between the original and surrogate series will differ. Because the distribution of the symbol statistics for the shuffled surrogates is unknown, one means of establishing confidence limits is to compute symbol statistics for a large number, say 200, of shuffled surrogates. Thus, one can mark the limits in which 95 percent of the statistics for the surrogate series lie. This method is a combination of Monte Carlo and bootstrapping techniques [11] and does not rely on traditional statistical inferencing methods which assume a certain (Gaussian) distribution.

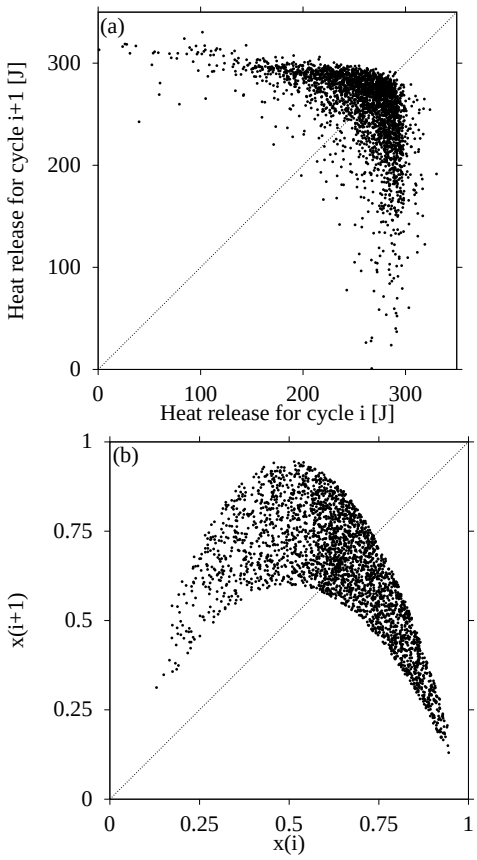

Figure 4: Lag-1 return maps for V8 data at equivalence ratio 0.63 (a) and a noisy period-2 logistic map (b). A simple binary partition does not distinguish the two data sets (see Fig. 5).

In **Fig. 2**, the mean relative frequency of all surrogate series for each bin (unique symbol sequence) is near the expected equiprobable value (1/64, or 0.015625). The confidence limits denote where 95 percent of all surrogate values were. Clearly, many of the significant peaks in the histogram for the original time series appear to be significantly greater than what is observed in the shuffled surrogates. This difference suggests with a high degree of confidence that there are indeed real patterns in the original time series. In **Fig. 3**, the curve for the original time series differs significantly from the shuffled-surrogate curve and corresponding confidence band.

Careful selection of symbol-set size must be exercised in order to highlight the proper features in a symbol series. For instance, as seen in **Fig. 2**, whereas $n = 2$ is quite sufficient for detecting the onset of the 0-1 alternating oscillations, we find that higher-level symbolization is needed to discriminate alternative models. A vivid example is illustrated in **Figs. 4** and **5**. The **Fig. 4a** is first return map for experimental engine heat-release data, whereas **Fig. 4b** is a first return map for the logistic map with Gaussian noise added to the feedback parameter. The logistic map is a form of a quadratic map and is related as $x_{i+1} = \lambda x_i(1 - x_i)$, where λ is a parameter. First return maps are formed by plotting each data point X_i against its immediately succeeding point X_{i+1}. For the logistic-map data, the mean value of the feedback parameter has been adjusted to produce a period-2 oscillation with no noise added. These return maps clearly reveal that the shapes of the underlying deterministic maps are different. Nevertheless, as illustrated in **Fig. 5a**, the symbol-sequence histograms are nearly identical. It would be very difficult to distinguish the two data sets (and thus determine the adequacy of the logistic map as a model) on the basis of these histograms. **Figure 5b** illustrates how much easier it is to distinguish the engine data from a noisy logistic map by increasing n from 2 to 4. Alternately, for $n = 2$, choice of partition other than the data median would also help to distinguish the two data sets.

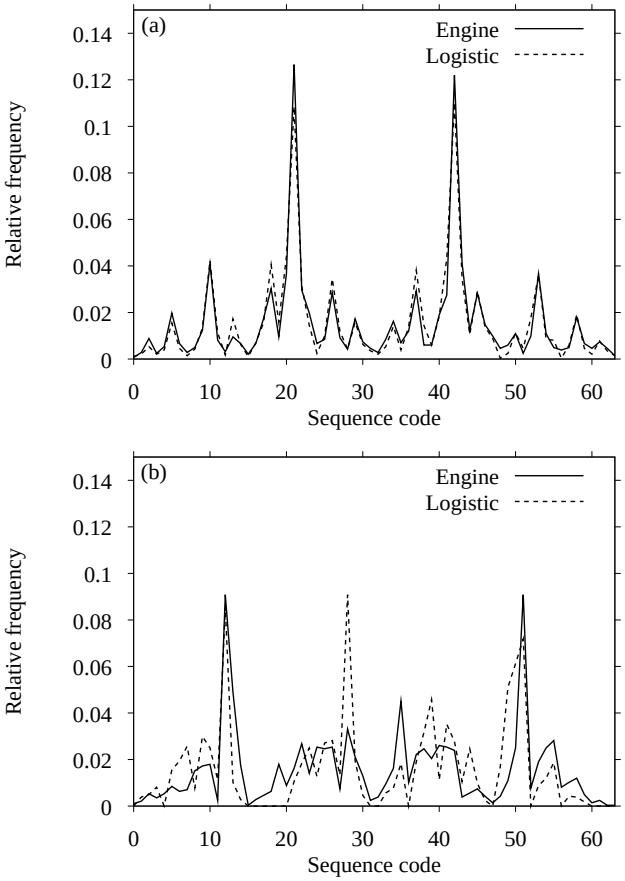

Figure 5: Symbol-sequence histograms for V8 data and a noisy period-2 logistic map. A binary partition ($n = 2$, $L = 6$), seen in (a), does not provide the discriminatory power seen with a quaternary partition ($n = 4$, $L = 3$), seen in (b).

All discussion hitherto has focused on the choice of a certain symbol-set size n to calculate symbol statistics for a given symbol-sequence length L. In some cases, it is not desirable to have the same n at each member in the symbol sequence. For instance, in the cases of fitting model to experimental data (see discussion below), we have found that using a different number of symbols in each step in the symbol sequence can provide better discrimination [3]. They sometimes provide better fits because the one-dimensional marginal distributions of trial simulations are constrained to match the data more closely when symbol partitions are not commensurate. Instead of using the parameters n and L to describe symbol-tree geometry, a more generic *symbolization code* made be used. This code is simply a list of symbol-set sizes at each element in the symbol sequence. For example, a symbolization code of $\{3\,2\}$ denotes 3 symbols for the first element of the symbol sequence and 2 symbols for the second element. Graphically, the symbol tree would be:

$$\begin{array}{ccccccc} & p_0 & & & p_1 & & \quad p_2 \\ p_{00} & & p_{01} & p_{10} & & p_{11} & p_{20} \quad p_{21} \end{array}$$

A practical example of a nonhomogeneous symbolization code will be shown below.

SYMBOLIC ANALYSIS OF COMBUSTION DATA

We focus here on discrete heat-release series, partly because of their utility in examination of the engine model described previously [1, 3]. However, we have found similar results for other measurements such as IMEP and cycle peak pressure.

BIFURCATION DETECTION — The model of Daw *et al.* [1] predicts that combustion variability should develop as a period-doubling bifurcation sequence as equivalence ratio is reduced from stoichiometric to lean. At stoichiometric fueling, combustion is expected to be regular and vary between cycles only slightly due to parametric noise; at sufficiently lean fueling, combustion is expected to oscillate between high and low heat release (the "period" doubling), with the extent of oscillation increasing as fueling becomes increasingly lean. In a noise-free experimental system, detection of the bifurcation would be straightforward. However, experimental systems are inherently noisy, from instrumentation and measurements as well as from dynamical effects such as cylinder-charge turbulence or fuel-atomization effects. Detection of the onset of bifurcations amidst measurement and parametric noise can be problematic.

Using symbol statistics, we demonstrate experimental observation of bifurcation in heat release in both experimental engines (this has also been observed in a third engine system [4]). **Figure 6** shows a series of symbol-sequence histograms of heat-release data from the V8 engine, and **Fig. 7** shows a series of symbol-sequence histograms of heat-release data from the single-cylinder engine. Both series of plots display the mean of 200 shuffled surrogates with 95 percent confidence limits. In the V8 series, for near-stoichiometric fueling, no predominant data patterns are visible in the histograms; as fueling becomes increasingly lean, peaks in the histograms at sequence codes 21, representing the symbol sequence 010101, and 42, representing the symbol sequence 101010, become significant. These patterns represent the onset and development of the bifurcation in heat release. Here, bifurcation refers to an increasing prevalence of alternating large and small heat releases which become symbolized as alternating 1 and 0. Note that with the behavior of the Daw model and with a binary symbolization, period-2 and higher periodicities (*e.g.*, period-4, -8 etc.) will all produce large frequencies of sequence codes 21 and 42.

In the single-cylinder data, similar patterns are visible. The general trend is that as equivalence ratio is reduced, the onset and development of a bifurcation is visible. In the first five plots (**Fig. 7a–e**), there is noticeable, real bias in symbol sequences 000000 (sequence code 0) and 111111 (63), as well as in the related transition sequences 010000 (16), 100000 (32) and 101111 (47). These sequences appear to occur because of transient or nonstationary dynamics in the engine, where the data mean drifts and data are symbolized as long sequences of 0 or 1. In these single-cylinder data, if the transients were not present, then the first five plots would be expected to resemble more closely the first four or five plots for the V8 data (**Fig. 6**).

The bifurcation is also visible in the modified Shannon entropy. At near-stoichiometric fueling, where combustion is expected to reflect a fairly steady (period-1) pattern but with some degree of noise, H_S is expected to be near 1, as the random effects of the noise are the most significant features in the data. At the onset and progress through the point of bifurcation, H_S is expected to decrease, as the detection of period-2 oscillations adds structure to the data, even if obscured by noise. **Figure 8** verifies these expectations for the V8 data shown in **Fig. 6**. At the point at which the bifurcation becomes significant, as seen in the symbol-sequence histograms, H_S decreases noticeably and significantly from random behavior.

CHARACTERIZATION OF TRANSIENTS — Engine combustion is very often transient in nature. An important and relevant example is cold start. On average, the mean heat release increases as the engine warms, so the detection of this transient is trivial. However, if the objective is to detect changes in the dynamical patterns (*e.g.*, increased frequency of misfires), symbolic analysis would be more useful. As noted previously, though, equiprobable partitioning is difficult in a transient context. One approach is to use a moving-window symbolization.

We focus here on identifying transients in which the engine dynamics appear to be stationary but undergo brief periods of bifurcated behavior. For instance, in a near-stoichiometric (unbifurcated) condition, the effects of dynamical noise can shift the engine into a bifurcated state for a brief period before the dynamics return to the nominal condition. **Figure 9** represents spectrograms of such a case. In these plots, symbol-sequence histograms are calculated in 100-point windows stepping every 10 points. The dark bands seen between cycles 20 and 50 and again between cycles 1030 and 1375 represent occurrences of period-2 behavior, as the dark shading in these two bands correspond to sequence codes 21 (sequence 010101) and 42 (sequence 101010). Also apparent is some nonstationarity between cycles 1500 and 1680, seen in the dark shading for sequence code 0 (sequence 000000). At most other times, sequence probabilities are relatively equiprobable (given the small number of points in the moving window), indicative of an apparent random relationship between heat release in successive cycles.

Figure 9b confirms these observations. Here, only symbol sequences which are significant with 95 percent confidence are displayed. Reducing the background clutter of **Fig. 9a** helps to identify which portions of the time record lead to significant fea-

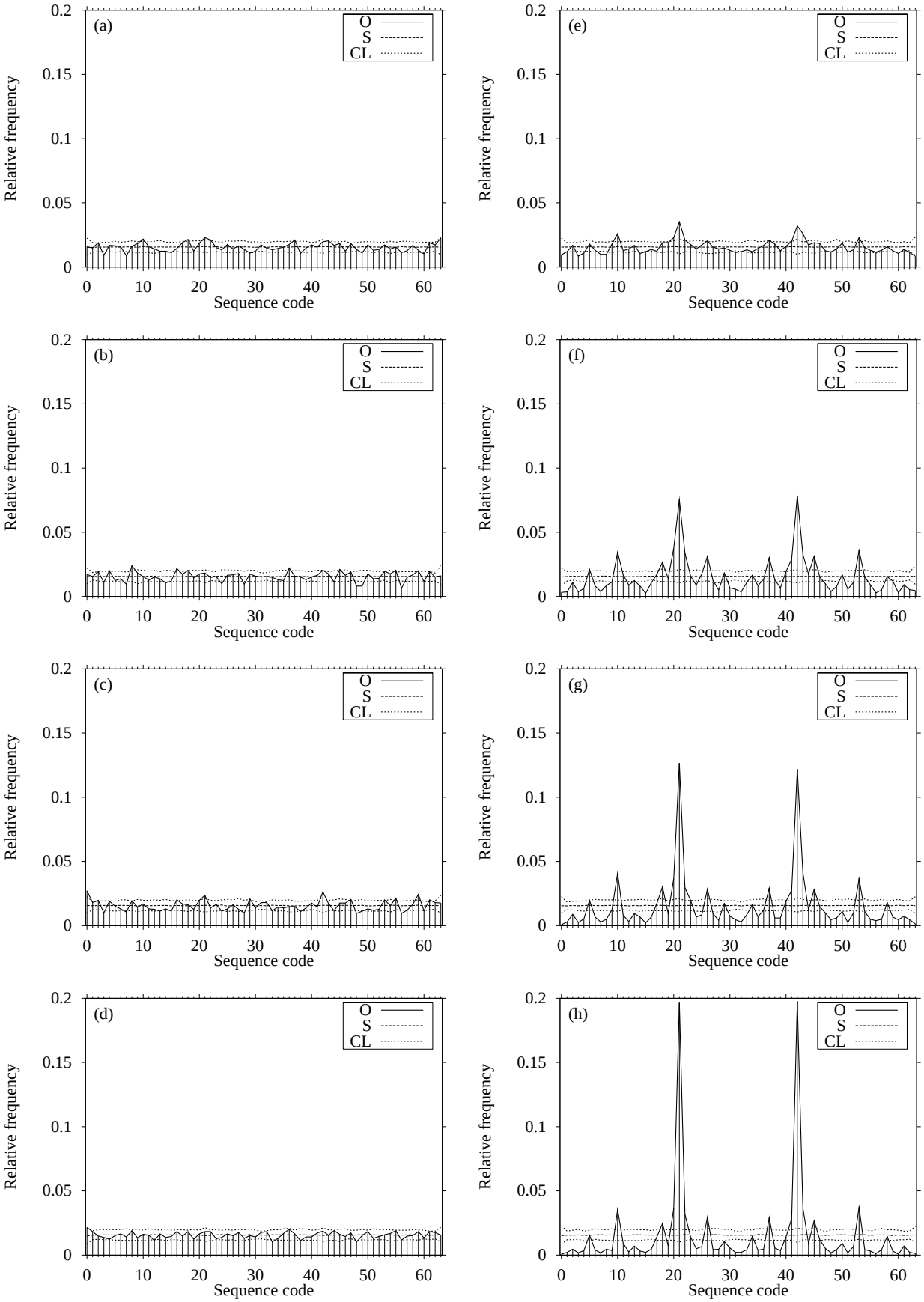

Figure 6: Symbol-sequence histograms for the V8 engine as equivalence ratio decreases from stoichiometric to moderately lean fueling. Solid lines ("O") represents the original symbol series, and dashed lines represent the mean ("S") and 95 percent confidence limits ("CL") for 200 shuffled surrogate series. Nominal equivalence ratios are 1.00 (a), 0.91 (b), 0.83 (c) 0.77 (d), 0.71 (e), 0.67 (f), 0.63 (g) and 0.59 (h).

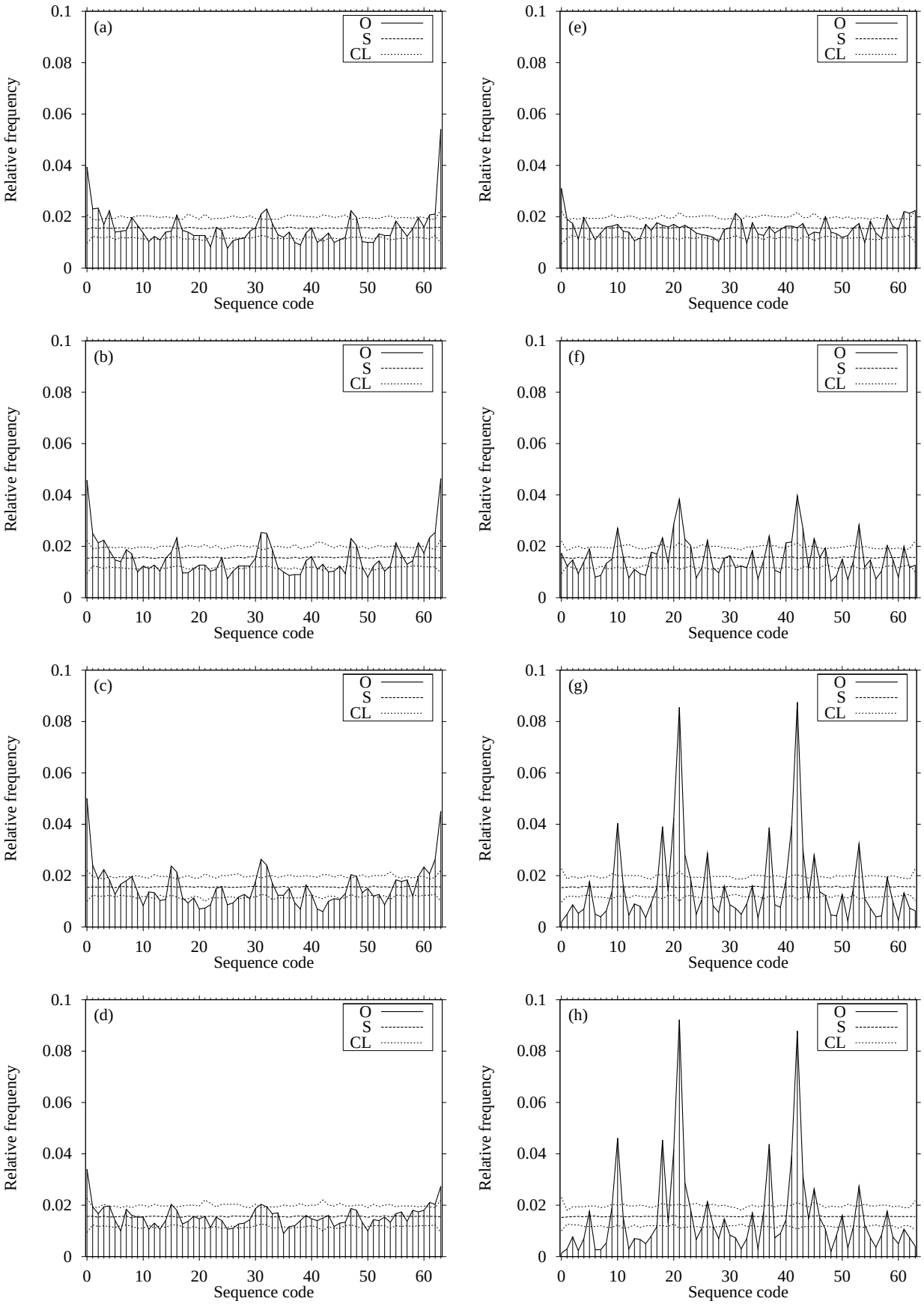

Figure 7: Symbol-sequence histograms for the single-cylinder engine as equivalence ratio decreases from stoichiometric to moderately lean fueling. Solid lines ("O") represents the original symbol series, and dashed lines represent the mean ("S") and 95 percent confidence limits ("CL") for 200 shuffled surrogate series. Nominal equivalence ratios are 0.99 (a), 0.96 (b), 0.84 (c) 0.80 (d), 0.77 (e), 0.75 (f), 0.73 (g) and 0.72 (h).

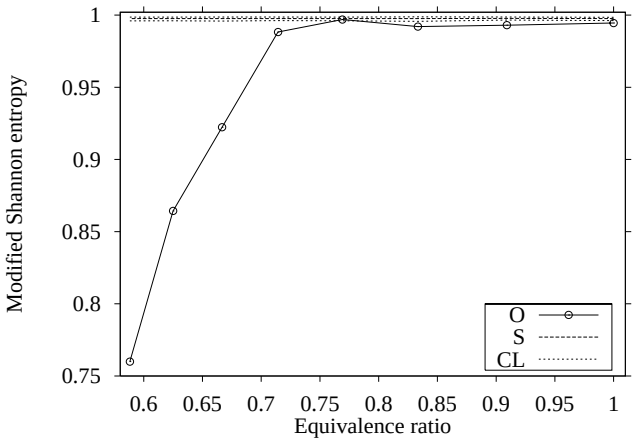

Figure 8: Modified Shannon entropy for the V8 engine as equivalence ratio decreases from stoichiometric to moderately lean fueling. Solid lines ("O") represents the original symbol series, and dashed lines represent the mean ("S") and 95 percent confidence limits ("CL") for 200 shuffled surrogate series.

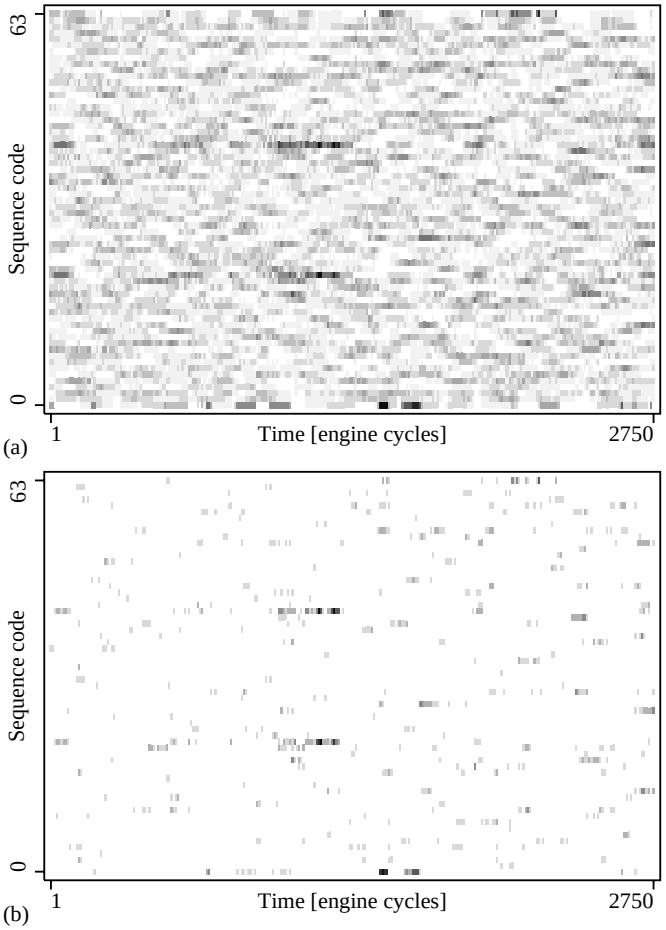

Figure 9: Symbol-sequence spectrograms for the V8 engine data at a nominal equivalence ratio of 0.83. Plot (a) represents the raw spectrogram, whereas plot (b) represents a filtered spectrogram in which only sequences with more than 95 percent confidence are displayed. The dark bands midway through the symbol series represent increased frequencies of sequences 010101 (sequence code 21) and 101010 (42), which suggests a brief period of bifurcated engine behavior.

tures of the overall symbol-sequence histogram (**Fig. 6c**). For instance, in **Fig. 6c** the sequences 000000, 010101 and 101010 are statistically significant, and the brief periods (listed above) where these patterns are dominant are evident in the spectrograms.

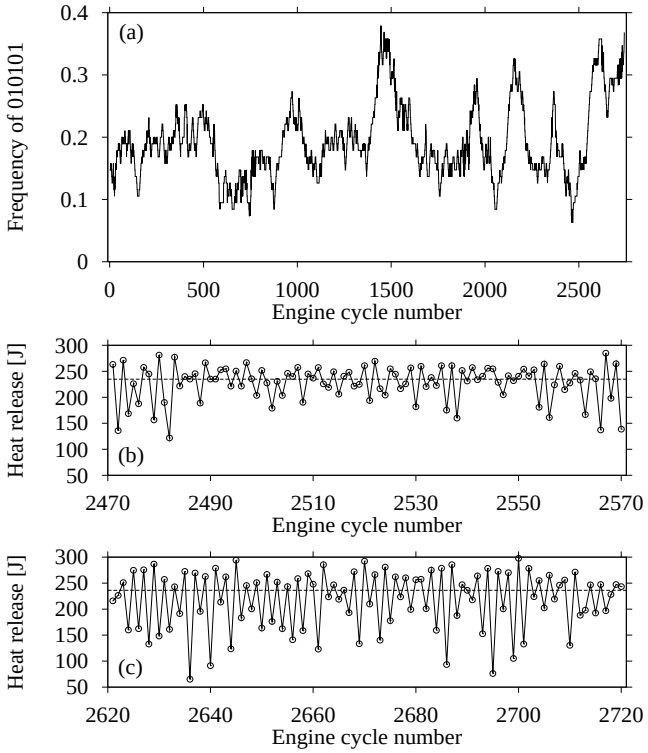

Figure 10: Symbol statistics and time-series segments for V8 engine data at a nominal equivalence ratio of 0.59. The strength of the 0-1 oscillation varies throughout the symbol series, as is evident in the change in frequency of symbol sequence 010101 (a). The time-series window in (b) has a low frequency of 010101, seen as a trough in (a) at cycle 2470, whereas the time-series window in (c) has a high frequency of 010101, seen as a peak in (a) at cycle 2620.

Additionally, such a moving-window symbol-sequence analysis allows the determination of periods of either weak or strong bifurcated behavior. For cases in which the desired pattern is known, tracking the relative frequency of only one or two symbol sequences might be a more desirable method of displaying this shift in behavior within a symbol series. As is visible in **Fig. 10**, the strength of the period-2 patterns exhibited in sequence code 21 varies throughout the symbol series. **Figure 10a** shows the relative frequency of symbol sequence 010101 calculated in a 100-point moving window over the entire time series, and **Fig. 10b–c** show the actual 100-point time-series windows with symbol partition which yield either low and high appearances of the symbol sequence 010101.

MODEL FITTING — In order to demonstrate that the model of Daw *et al.* [1] produces simulated engine behavior which mimics the dynamics of an experimental engine, we employed a genetic-algorithm-type search technique to compare symbol-sequence histograms of experimental and model data [3]. The use of symbol-sequence histograms guaranteed that not only did experimental and model data appear similar

(have similar data distributions) over a range of operating conditions, but also that they exhibited similar dynamical behavior. As noted above, a binary partition did not provide adequate discrimination to compare an experimental test data set with trial model data sets. Instead, a more-refined partitioning scheme was used to define in better detail the dynamical patterns present in both the test and trial data sets.

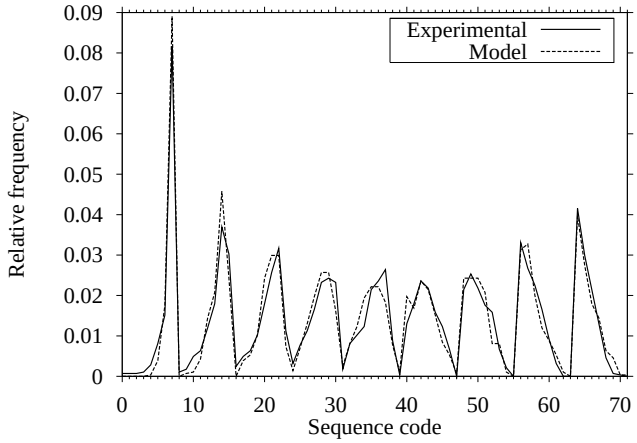

Figure 11: Symbol-sequence histograms using a symbolization code of $\{9\ 8\}$ for V8 experimental and model data at a nominal equivalence ratio of 0.59.

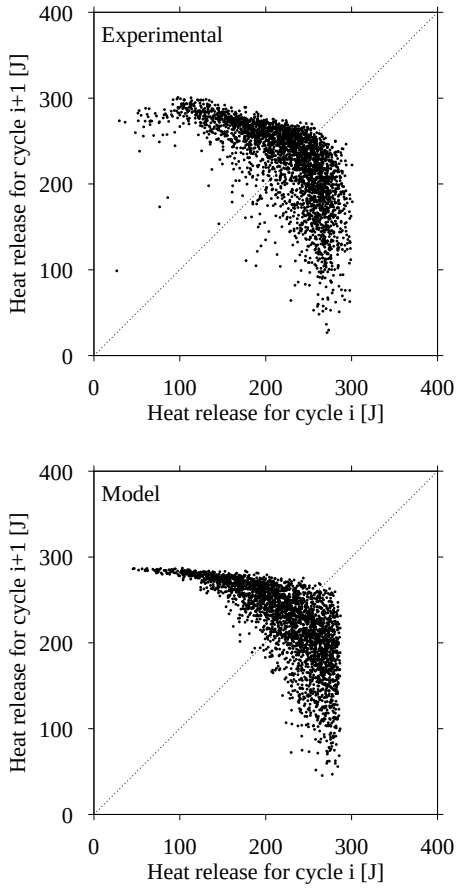

Figure 12: First return maps of V8 experimental and model data at a nominal equivalence ratio of 0.59.

Figure 11 compares symbol-sequence histograms using a symbolization code of $\{9\ 8\}$ for V8 experimental and model data at a nominal equivalence ratio of 0.59 (note that $\{9\ 8\}$ refers to a symbol-set size of 9 for the first element of the symbol sequence and a symbol-set size of 8 for the second element). The nature of the histogram is less symmetric than was the case with binary symbolization, and the degree of similarity for most patterns (sequence codes) suggests dynamical similarity between the experimental and model data. **Figure 12** displays the return maps for the experimental and fitted model data. There is good geometric similarity except for systematic differences seen in the upper-left portion of the experimental data which we have attributed to deficiencies in the model. This example, however, shows that the use of symbol statistics can provide a comparison of dynamical features from different data sets, and the insensitivity to noise of symbolic analysis is especially useful for analyzing real, experimental data.

SUMMARY AND CONCLUSIONS

We have presented methods of symbolic analysis of dynamic engine combustion measurements and shown their utility in detecting deterministic features in data amidst high levels of noise. We have shown that symbolic techniques can be used to characterize the nature of transient patterns in data. Because of their computational speed and insensitivity to noise, data symbolization and symbolic analysis show promise for implementation in real-time experimental, and more significantly, production engines for detecting bifurcation and transient dynamics. We also anticipate that these methods are applicable in non-combustion measurements (*e.g.*, vibration).

NOMENCLATURE

H_S	Modified Shannon entropy
L	Symbol-sequence length
n	Symbol-set size
N_{seq}	Number of non-zero-frequency sequences
$p_{i,L}$	Probability of sequence i of symbol-sequence length L
$\{n_1 \ldots n_L\}$	Symbolization code relating the symbol-set size at each step in a symbol sequence

ACKNOWLEDGEMENTS

The authors thank F.T. Connolly of the Ford Motor Company for supplying experimental data from the V8 system and J.S. Armfield of the Oak Ridge National Laboratory for experimental-apparatus development of the single-cylinder system.

REFERENCES

[1] C.S. Daw, C.E.A. Finney, J.B. Green, Jr., M.B. Kennel, J.F. Thomas, and F.T. Connolly, "A simple model for cyclic variations in a spark-ignition engine", SAE Paper No. 962086, 1996.

[2] C. Letellier, S. Meunier-Guttin-Cluzel, G. Gouesbet, F. Neveu, T. Duverger, and B. Cousyn, "Use of the nonlin-

ear dynamical system theory to study cycle-to-cycle variations from spark-ignition engine pressure data", SAE Paper No. 971640, 1997.

[3] C.S. Daw, C.E.A. Finney, M.B. Kennel, and F.T. Connolly, "Cycle-by-cycle variations in spark-ignited engines", *Proceedings of the Fourth Experimental Chaos Conference*, 1997.

[4] R.M. Wagner, J.A. Drallmeier, C.S. Daw, "Prior-cycle effects in lean spark ignition combustion: fuel/air charge considerations", SAE Paper No. 981047, to be published, 1998.

[5] J.B. Heywood. **Internal Combustion Engine Fundamentals**, McGraw-Hill, ISBN 0-07-028637-X, 1988.

[6] B.M. Grimm and R.T. Johnson, "Review of simple heat release computations", SAE Paper No. 900445, 1990.

[7] X.Z. Tang, E.R. Tracy, A.D. Boozer, A. deBrauw, and R. Brown, "Symbol sequence statistics in noisy chaotic signal reconstruction", *Physical Review E* **51**:5, 3871–3889, 1995.

[8] M. Lehrman, A.B. Rechester, and R.B. White, "Symbolic analysis of chaotic signals and turbulent fluctuations", *Physical Review Letters* **78**:1, 54–57, 1997.

[9] J.P. Crutchfield and N.H. Packard, "Symbolic dynamics of noisy chaos", *Physica D* **7**, 201–223, 1983.

[10] C. Diks, J.C. van Houwelingen, F. Takens and J. DeGoede, "Reversibility as a criterion for discriminating time series", *Physics Letters A* **201**, 221–228, 1995.

[11] J. Theiler and D. Prichard, "Constrained-realization Monte-Carlo method for hypothesis testing", *Physica D* **94**, 221–235, 1996.