Detecting Edgeworth Cycles*

Timothy Holt† Mitsuru Igami‡ Simon Scheidegger§

November 18, 2021

Abstract

We propose algorithms to detect "Edgeworth cycles," asymmetric price movements that have caused antitrust concerns in many countries. We formalize four existing methods and propose six new methods based on spectral analysis and machine learning. We evaluate their accuracy in station-level gasoline-price data from Western Australia, New South Wales, and Germany. Most methods achieve high accuracy in the first two, but only a few can detect nuanced cycles in the third. Results suggest whether researchers find a positive or negative statistical relationship between cycles and markups, and hence their implications for competition policy, crucially depends on the choice of methods.

Keywords: Edgeworth cycles, Fuel prices, Markups, Nonparametric methods

JEL classifications: C45 (Neural Networks and Related Topics), C55 (Large Data Sets: Modeling and Analysis), L13 (Oligopoly and Other Imperfect Markets), L41 (Monopolization, Horizontal Anticompetitive Practices).

*We thank Xiaohong Chen for advice on time-series methods, Robert Clark, Daniel Ershov, and Simon Martin for advice on the German data, Kevin Schäfer and the Argus Media group for kindly providing us with the German wholesale prices, and our team of research assistants at Yale University (Yue Qi, Alan Chiang, Alexis Teh, Clara Penteado, Janie Wu, Bruno Moscarini, Eileen Yang, and Jordan Mazza) for excellent work. We also thank Roxana Mihet and seminar participants at the Düsseldorf Institute for Competition Economics (DICE), Brown University, Cornell University, and Yale University for comments. This work was supported by the Swiss National Science Foundation (SNF), under project ID "New methods for asset pricing with frictions."

†Institute of Computing, Università della Svizzera italiana. E-mail: timothy.holt@usi.ch.

‡Yale Department of Economics. E-mail: mitsuru.igami@yale.edu.

§Department of Economics, HEC Lausanne. E-mail: simon.scheidegger@unil.ch.

1 Introduction

Retail gasoline prices are known to follow cyclical patterns in many countries (e.g., Byrne and de Roos 2019). The patterns persist even after controlling for the wholesale and crude-oil prices. Because these cycles are so regular and conspicuous, and because price increases tend to be larger than decreases, observers suspect anti-competitive business practices. The occasional discovery of price-fixing cases supports this view (e.g., Clark and Houde 2014, Foros and Steen 2013, Wang 2008). Moreover, recent studies on algorithmic collusion suggest interactions between self-learning algorithms could lead to collusive equilibria with such cycles (Klein 2021); the use of "repricing algorithms" by many sellers on Amazon has made these phenomena prevalent in e-commerce as well (Musolff 2021).

These asymmetric movements are called Edgeworth cycles and have been studied extensively.[1] In particular, scholars and antitrust practitioners have investigated whether the presence of cycles is associated with higher prices and markups. Deltas (2008), Clark and Houde (2014), and Byrne (2019) find asymmetry is correlated with higher margins, price-fixing collusion, and concentrated market structure, respectively. However, Lewis (2009), Zimmerman et al. (2013), and Noel (2015) show prices and margins are *lower* in markets with asymmetric price cycles. Given the diversity of countries and regions in these studies (Australia, Canada, the US, and several countries in Europe), the cycle-competition relationship could be intrinsically heterogeneous across markets.

But another, perhaps more fundamental problem is measurement: the lack of a formal definition or a reliable method to detect cycles in large datasets. Some researchers rely on visual inspections alone; others use more objective criteria, such as the sign of the median daily change.[2] Unfortunately, most of these operational definitions focus exclusively on asymmetry and do not capture cyclicality. Even though asymmetry may be the most salient feature of—and hence a necessary condition for—Edgeworth cycles, it is not a sufficient condition. Because empirical findings are functions of their definition, these issues directly affect our ability to establish "facts" about competition and price cycles. Now that the governments of many countries and regions are making large-scale price data publicly available,[3] developing

[1] Maskin and Tirole (1988) coined the term after Edgeworth's (1925) hypothetical example. It became a popular topic for empirical research since Castanias and Johnson (1993).

[2] The negative median change could indicate Edgeworth cycles, because they feature many days of small price decreases and occasional large price increases.

[3] The governments of Australia, Germany, and other countries have made detailed price data publicly available to inform consumers and encourage further scrutiny. The Australian Consumer and Competition Commission has a team dedicated to monitoring gasoline prices and regularly publishes reports. See https://www.accc.gov.au/consumers/petrol-diesel-lpg/about-fuel-prices.

a scalable detection method represents an important practical challenge for economists and policymakers.

This paper proposes a systematic approach to detecting Edgeworth cycles. We formalize four existing methods as simple parametric models: (1) the "positive runs vs. negative runs" method of Castanias and Johnson (1993), (2) the "mean increase vs. mean decrease" method of Eckert (2002), (3) the "negative median change" method of Lewis (2009), and (4) the "many big price increases" method of Byrne and de Roos (2019). We then propose six new methods based on spectral analysis and nonparametric/machine-learning techniques: (5) Fourier transform, (6) the Lomb-Scargle periodogram, (7) cubic splines, (8) long short-term memory (LSTM), (9) an "ensemble" (aggregation) of Methods 1–7 within a random-forests framework, and (10) an ensemble of Methods 1–8 within an extended LSTM.[4]

To evaluate the performance of each method, we collect data on retail and wholesale gasoline prices in two regions of Australia, Western Australia (WA) and New South Wales (NSW), as well as the entirety of Germany. These datasets cover the universe of gas stations in these regions/countries, record each station's retail price at a daily (or higher) frequency, and are made publicly available by legal mandates.[5] We construct a benchmark "ground truth" based on human recognition of price cycles as follows. We reorganize the raw data as panel data of gas stations' daily prices and group them into calendar quarters, so that a station-quarter (i.e., a set of 90 consecutive days of retail-margin observations for each station) becomes the effective unit of observation. We train research assistants (RAs) to manually classify each station-quarter as either "cycling," "maybe cycling," or "not cycling." We then define a binary indicator variable that equals 1 if an observation is labeled as "cycling" by all of the RAs (the majority of observations are labeled by three RAs), and 0 otherwise, thereby preparing a conservative target for automatic cycle detection.[6]

We report three sets of results. First, when applied to the two Australian datasets, most of the methods—both existing and new—achieve high accuracy levels near or above 90% and 80%, respectively, because price cycles are clearly asymmetric and exhibit regular periodicity (hence, easy to detect) in these regions. By contrast, German cycles are more subtle and diverse, defying many methods. All existing methods except Method 4 fail to detect cycles, even though as much as 40% of the sample is unanimously labeled as "cycling" by three RAs (see Figure 1 for examples). This failure is not an artifact of sample selection or human error,

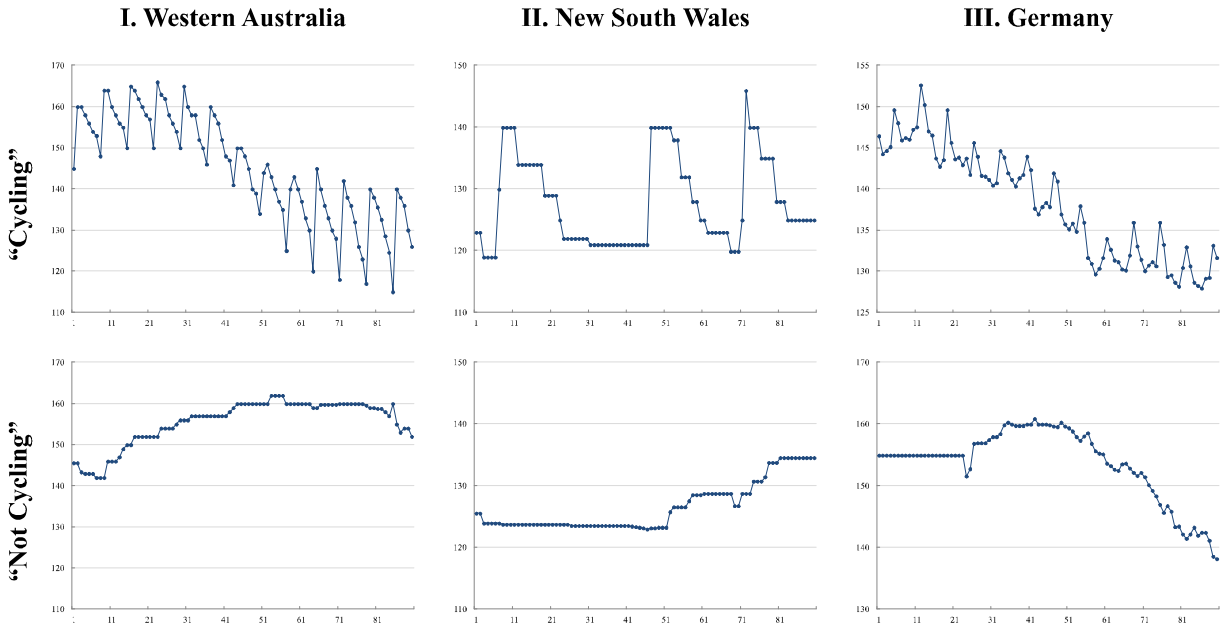
[4]Section 3 formally introduces all methods.

[5]See Byrne, Nah, and Xue (2018) for a guide to the Australian data. Haucap, Heimeshoff, and Siekmann (2015), Martin (2018), and Assad, Clark, Ershov, and Xu (2021), among others, study the German data.

[6]Section 2 explains the data and the classification procedures. Appendix C shows results under an alternative criterion.

because our interview with a German industry expert suggests Edgeworth cycles are known to exist, called the "price parachute" (or *Preis Fallschirm*) phenomena, and considered to be part of common pricing strategies among practitioners. Methods 7–10 attain 71%–80% accuracy even in this challenging environment.

Figure 1: Examples of Cycling and Non-cycling Station-Quarter Observations



Note: The top panels and the bottom panels are examples of daily retail-price series in "cycling" and "non-cycling" station-quarter observations, respectively, for illustration purposes. The vertical axes measure retail gas prices in the Australian cent (left, middle) and the Euro cent (right), respectively. The horizontal axes represent calendar days. See section 2 and Appendix A for our manual classification procedures, and section 3 for the automatic methods.

Second, we assess the cost effectiveness of each method by using only 0.1%, 1%, 5%, and so on of our manually labeled subsamples as "training" data. Results suggest simpler models (Methods 1–7) are extremely "cheap" to train, as they quickly approach their respective maximal accuracy with only a dozen observations. The nonparametric models (Methods 8–10) need more data to achieve near-maximal performances, but their data requirement is still sufficiently small for practical purposes. Only a few hundred observations prove sufficient for even the most complex model (Method 10). The economic cost of manually classifying a few hundred observations is in the order of tens of RA hours, or a few hundred US dollars at the current hourly wage of US$13.50 for undergraduate RA work at Yale University. Potential cost savings are sizable, as manually labeling the entire German dataset in 2014–2020 would

require 4,800 RA hours, or US$64,800. Thus, our approach is economical and suitable for researchers and governments with limited resources.

Third, we investigate whether and how gas stations' markups are correlated with the presence of cycles. In WA and NSW, the average margins in (manually classified) "cycling" station-quarters are statistically significantly higher than in "non-cycling" ones. The relationship is reversed in Germany, where the margins in "cycling" observations are lower than in "non-cycling" ones. Hence, in general, the presence of cycles could be either positively or negatively correlated with markups. All of the automatic detection methods lead to the correct finding (i.e., positive correlations) in WA, but some of them fail in NSW. Furthermore, Methods 1–6 either fail to detect cycles or lead to false conclusions in Germany (i.e., find statistically significant *positive* correlations). Thus, whether researchers discover a positive, negative, or no statistical relationship between markups and cycles—a piece of highly policy-relevant empirical evidence—depends on the seemingly innocuous choice of operational definitions. Appendix E summarizes our methodological recommendations as a "practitioner's guide."

The rest of the paper is organized as follows. Sections 2 and 3 explain the data and the methods, respectively. Sections 4–6 report results and discuss their economic/policy implications. Section 7 concludes.

2 Data and Manual Classification

Retail-price data are publicly available for the universe of individual gas stations in WA, NSW, and Germany.[7] We combine them with wholesale-price data, based on the region of each station (Australia) or the location of the nearest refinery (Germany). We compute station-level daily profit margins by subtracting the relevant wholesale price from retail price, and simply refer to this measure as "price" in the following. We organize these daily prices by calendar quarter, so that station-quarter (i.e., a sequence of daily prices over 90 days for each station) becomes the unit of observation for cycle detection.[8]

To establish a "ground truth" based on human recognition of cycles, we trained a team

[7] See Appendix A.1 for the details on the data sources.

[8] Our choice of the daily sampling frequency and the quarterly time window reflects the following considerations. On the one hand, shorter-than-daily frequencies are irrelevant in WA (Appendix A.1 explains why), seem to increase the noise in the NSW data, and tend to pick up the intra-day "dynamic pricing" fluctuations in Germany. On the other hand, macroeconomic shocks tend to feature prominently in longer-than-quarterly time horizons, which are not our main subject either.

of eight RAs to manually classify station-quarter observations.[9] Each station i in quarter t is classified as either "cycling," "maybe cycling," or "not cycling." The total number of manually labeled observations is 24,569 (WA), 9,693 (NSW), and 35,685 (Germany). The RAs' total working hours are approximately 260 (WA), 210 (NSW), and 480 (Germany).[10]

Table 1: Summary Statistics

	(1)	(2)	(3)
Dataset	Western Australia	New South Wales	Germany
Sample period (yyyy/mm/dd)	$2001/1/3 - 2020/6/30$	$2016/8/1 - 2020/7/31$	$2014/6/8 - 2020/1/7$
Number of gasoline stations	821	1,226	14,780
Number of calendar quarters	77	15	26
Number of station-quarters	25,463	9,693	353,086
Of which:			
Labeled as "cycling" by 3 RAs	0 (0.0%)	6,878 (71.0%)	14,116 (39.6%)
Labeled as "cycling" by 2 RAs	0 (0.0%)	906 (9.4%)	7,173 (20.1%)
Labeled as "cycling" by 1 RA	15,007 (61.1%)	759 (7.8%)	6,280 (17.6%)
Not labeled as "cycling" by any RA	9,562 (38.9%)	1,150 (11.9%)	8,116 (22.7%)
Total manually labeled	24,569 (100.0%)	9,693 (100.0%)	35,685 (100.0%)
Not manually labeled	894	0	317,401

Note: Each "manually labeled" station-quarter observation in the WA data is single-labeled as either "cycling," "maybe cycling," or "not cycling," whereas the NSW and German data are triple-labeled. See Appendix A for details.

Table 1 reports summary statistics. Based on these manual-classification results, we define $cycle_{i,t}$ as a binary variable indicating the presence of clear cycles. In WA, each observation is labeled exactly once, based on the consensus of two RAs. We set $cycle_{i,t} = 1$ if station-quarter (i, t) is labeled as "cycling," and 0 otherwise. In the NSW and German data, which contain more ambiguous patterns, we assigned three RAs to label each observation individually, and hence each (i, t) is triple-labeled. We set $cycle_{i,t} = 1$ for observations with triple "cycling" labels (i.e., based on three RAs' unanimous decision), and 0 otherwise.[11] Thus, we prepare the target for automatic detection in a relatively conservative manner.

3 Models and Methods for Automatic Detection

This section explains (i) how we formalize the four existing methods, (ii) the six new methods that we propose, and (iii) the way we optimize the parameter values of each model.

[9] All of them are graduate or undergraduate students majoring in economics, mathematics, and statistics at Yale University.

[10] See Appendix A.2 for further details on the manual labeling procedures.

[11] See Appendix C for results based on an alternative criterion.

3.1 Existing Methods

The existing methods in the literature almost exclusively focus on asymmetric cycles. We formalize four of them as simple parametric models.[12]

Method 1: Positive Runs vs. Negative Runs ("PRNR"). Castanias and Johnson (1993) compare the lengths of positive and negative changes. We formalize this idea by classifying each station-quarter as cycling ($cycle_{i,t} = 1$) if and only if

$$mean\left(len\left(run^{+}\right)\right) < mean\left(len\left(run^{-}\right)\right) + \theta^{PRNR}, \tag{1}$$

where $len\left(run^{+}\right)$ and $len\left(run^{-}\right)$ denote the lengths of consecutive (multi-day) price increases/zero changes and decreases within quarter t, respectively. The means are taken over these "runs." $\theta^{PRNR} \approx 0$ is a scalar threshold, which we treat as a parameter.[13]

Method 2: Mean Increase vs. Mean Decrease ("MIMD"). Eckert (2002) compares the magnitude of the mean increase and the mean decrease. Formally, station-quarter (i, t) is cycling if and only if

$$\left|mean_{d\in t}\left(\Delta p_{i,d}^{+}\right)\right| > \left|mean_{d\in t}\left(\Delta p_{i,d}^{-}\right)\right| + \theta^{MIMD}, \tag{2}$$

where $\Delta p_{i,d}^{+}$ and $\Delta p_{i,d}^{-}$ denote positive and negative daily price changes at station i (between days d and $d-1$), respectively, and $\theta^{MIMD} \approx 0$ is a scalar threshold. That is, a cycle is detected when the average price increase is greater than the average price decrease.

Method 3: Negative Median Change ("NMC"). Lewis (2009) and many subsequent papers classify $cycle_{i,t} = 1$ if and only if

$$median_{d\in t}\left(\Delta p_{i,d}\right) < \theta^{NMC}, \tag{3}$$

where $\Delta p_{i,d}$ denotes price change between days d and $d-1$, and $\theta^{NMC} \approx 0$ is a scalar threshold. In other words, the significantly negative median change is taken as evidence of price cycles.

[12] See Appendix B.1 for a review of other methods and papers.

[13] Eckert (2002) proposes a more comprehensive version of this idea, which compares the *distributions* of positive and negative runs across lengths, by using the Kolmogorov-Smirnov test.

Method 4: Many Big Price Increases ("MBPI"). Byrne and de Roos (2019) identify price cycles with the condition

$$\sum_{d \in t} \mathbb{I}\left\{\Delta p_{i,d} > \theta_1^{MBPI}\right\} \geq \theta_2^{MBPI}, \tag{4}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function that equals 1 if the condition inside the bracket is satisfied, and 0 otherwise. θ_1^{MBPI} and θ_2^{MBPI} are thresholds for "big" and "many" price increases, respectively. They set $\theta_1^{MBPI} = 6$ (Australian cents/liter) and $\theta_2^{MBPI} = 3.75$ (per quarter) in studying the WA data. Thus, many instances of big price increases are taken as evidence of price cycles.

3.2 Our Proposals

We propose six new methods. Methods 5–6 are based on spectral analysis, and hence are attractive as formal mathematical definitions of regular cycles. By contrast, Methods 7–8 build on nonparametric regressions and machine-learning techniques, respectively, and are more suitable for capturing nuanced patterns and replicating human recognition of cycles. Methods 9–10 combine some or all of the previous methods. Appendix B.2 provides additional technical details.

Method 5: Fourier Transform ("FT"). Fourier analysis is a mathematical method for detecting and characterizing periodicity in time-series data. When a continuous function of time $g(x)$ is sampled at regular time intervals with spacing Δx, the sample analog of the Fourier power spectrum (or "periodogram") is

$$P(f) \equiv \frac{1}{N}\left|\sum_{n=1}^{N} g_n e^{-2\pi i f x_n}\right|^2, \tag{5}$$

where f is frequency, N is the sample size, $g_n \equiv g(n\Delta x)$, $i \equiv \sqrt{-1}$ is the imaginary unit (not to be confused with our gas-station index), and x_n is the time stamp of the n-th observation. It is a positive, real-valued function that quantifies the contribution of each frequency f to the time-series data $(g_n)_{n=1}^N$.[14]

[14] Appendix B.2 introduces FT to the readers who are not familiar with Fourier analysis.

We focus on the highest point of $P(f)$ and detect cycles if and only if

$$\max_f P_{i,t}(f) > \theta^{FT}_{\max}, \tag{6}$$

where $P_{i,t}(f)$ is the periodogram (5) of station-quarter (i,t), and $\theta^{FT}_{\max} > 0$ is a scalar threshold parameter.

Method 6: Lomb-Scargle ("LS") Periodogram. The LS periodogram (Lomb 1976, Scargle 1982) characterizes periodicity in unevenly sampled time series.[15] It has been extensively used in astrophysics because astronomical observations are subject to weather conditions and diurnal, lunar, or seasonal cycles. Formally, it is a generalized version of the classical periodogram (5):[16]

$$P^{LS}(f) = \frac{1}{2}\left\{\frac{\left(\sum_n g_n \cos\left(2\pi f\left[x_n-\tau\right]\right)\right)^2}{\sum_n \cos^2\left(2\pi f\left[x_n-\tau\right]\right)} + \frac{\left(\sum_n g_n \sin\left(2\pi f\left[x_n-\tau\right]\right)\right)^2}{\sum_n \sin^2\left(2\pi f\left[x_n-\tau\right]\right)}\right\}, \tag{7}$$

where τ is specified for each frequency f as

$$\tau = \frac{1}{4\pi f}\tan^{-1}\left(\frac{\sum_n \sin\left(4\pi f x_n\right)}{\sum_n \cos\left(4\pi f x_n\right)}\right). \tag{8}$$

We propose the following threshold condition to detect cycles:

$$\max_f P^{LS}_{i,t}(f) > \theta^{LS}_{\max}, \tag{9}$$

where $\theta^{LS}_{\max} > 0$ is a scalar threshold parameter.

Method 7: Cubic Splines ("CS"). This method captures cycles' frequency in a less structured manner than FT and LS by using cubic splines. A spline is a piecewise polynomial function. We smooth the discrete (daily) time series by interpolating it with a cubic Hermite interpolater, which is a spline where each piece is a third-degree polynomial of Hermite form.[17] For each (i,t), we fit CS to its demeaned price series, $\overline{p}_{i,d} \equiv p_{i,d} - mean_{d\in t}(p_{i,d})$, and count the number of times the fitted function $\overline{CS}_{i,t}(d)$ crosses the d-axis (i.e., equals 0).

[15] Our data are evenly sampled at the daily frequency and can be analyzed by FT alone, but the LS periodogram offers additional benefits. One is conceptual: it is interpretable as a kind of nonparametric regression (see Appendix B.2). Another is practical: its off-the-shelf computational implementation can offer more granular periodograms.

[16] Appendix B.2 explains how this expression relates to FT.

[17] Appendix B.2 explains the details of this functional form.

That is, we count the number of real roots and detect cycles with the condition,

$$\#roots\left(\overline{CS}_{i,t}\left(d\right)\right) > \theta_{root}^{CS}, \tag{10}$$

where $\theta_{root}^{CS} > 0$ is a scalar parameter. Thus, any frequent oscillations (not limited to the sinusoidal ones as in FT or LS) become a sign of cycles.

Method 8: Long Short-Term Memory ("LSTM"). Recurrent neural networks with LSTM (Hochreiter and Schmidhuber 1997) are a class of artificial neural networks (ANNs) models for sequential data. LSTM networks have become a "de-facto standard" for recognizing and predicting complicated patterns in many applications, including speech, handwriting, language, and polyphonic music. Because LSTM is relatively new, we explain this method in greater detail.

Econometrically speaking, LSTM is a nonparametric model for time-series analysis. It is a recursive dynamic model whose behavior centers on a collection of pairs of $B_l \times 1$ vector-valued latent state variables, $\mathbf{s}_d^l$ and $\mathbf{c}_d^l$, where $l = 1, 2, \cdots, L$ is an index of layers. As this notation suggests, we are using a multi-layer architecture (a.k.a. "deep" neural networks) to enhance the model's flexibility.[18] B_l represents the number of blocks per layer, which are analogous to "neurons" (basic computing units) in other ANN models. $\mathbf{s}_d^l$ is an output state that represents the current, "short-term" state, whereas $\mathbf{c}_d^l$ is called a cell state and retains "long-term memory." The latter is designed to capture lagged dependence between the state and input variables, thereby playing the role of a memory cell in electronic computers.

These state variables evolve according to the following Markov process:

$$\mathbf{s}_d^l = \underbrace{\tanh\left(\mathbf{c}_d^l\right)}_{\text{"output"}} \circ \underbrace{\Lambda\left(\boldsymbol{\omega}_1^l + \boldsymbol{\omega}_2^l \Delta p_d + \boldsymbol{\omega}_3^l \mathbf{s}_d^{l-1}\right)}_{\text{"output gate"}}, \text{ and} \tag{11}$$

$$\begin{aligned}\mathbf{c}_d^l = {} & \underbrace{\tanh\left(\boldsymbol{\omega}_4^l + \boldsymbol{\omega}_5^l \Delta p_d + \boldsymbol{\omega}_6^l \mathbf{s}_d^{l-1}\right)}_{\text{"input"}} \circ \underbrace{\Lambda\left(\boldsymbol{\omega}_7^l + \boldsymbol{\omega}_8^l \Delta p_d + \boldsymbol{\omega}_9^l \mathbf{s}_d^{l-1}\right)}_{\text{"input gate"}} \\ & + \mathbf{c}_d^{l-1} \circ \underbrace{\left[1 - \Lambda\left(\boldsymbol{\omega}_7^l + \boldsymbol{\omega}_8^l \Delta p_d + \boldsymbol{\omega}_9^l \mathbf{s}_d^{l-1}\right)\right]}_{\text{"forget gate"}},\end{aligned} \tag{12}$$

where $d = 1, 2, \cdots, D$ is our index of days, $\Delta p_d \equiv p_d - p_{d-1}$ (we set $\Delta p_1 = 0$), $\tanh\left(x\right) \equiv \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is the hyperbolic tangent function, $\circ$ denotes the Hadamard (element-wise) product,

[18] Except for the multi-layer design, our specification mostly follows Greff et al. (2017), in which one of the original proponents of LSTM and his team compare many of its variants and show their simple "vanilla" specification outperforms others.

and $\Lambda(x) \equiv \frac{e^x}{1+e^x}$ is the cumulative density function (CDF) of the logistic distribution.[19] The $\boldsymbol{\omega}$s are weight parameters with the following dimensionality: (i) $\boldsymbol{\omega}_1^l$, $\boldsymbol{\omega}_2^l$, $\boldsymbol{\omega}_4^l$, $\boldsymbol{\omega}_5^l$, $\boldsymbol{\omega}_7^l$, and $\boldsymbol{\omega}_8^l$ are $B_l \times 1$ vectors; and (ii) $\boldsymbol{\omega}_3^l$, $\boldsymbol{\omega}_6^l$, and $\boldsymbol{\omega}_9^l$ are $B_l \times B_{l-1}$ matrices. Thus, $\mathbf{B} \equiv (B_1, B_2, \cdots, B_L)$ determines the effective number of latent state variables and parameters, and hence the flexibility of the model.

The first layer $l = 1$ of time d takes as input the states of the last layer $l = L$ of time $d-1$. Thus, $\left(\mathbf{s}_d^{l-1}, \mathbf{c}_d^{l-1}, B_{l-1}\right)$ in the above should be replaced by $\left(\mathbf{s}_{d-1}^{L}, \mathbf{c}_{d-1}^{L}, B_L\right)$ when $l = 1$. After the final layer L of the last day $D = 90$ of quarter t, we detect cycles in station-quarter (i, t) if and only if

$$s^*\left(\mathbf{p}_{i,t}; \boldsymbol{\theta}^{LSTM}\right) \equiv \omega_{10} + \boldsymbol{\omega}_{11}' \mathbf{s}_D^L > 0, \tag{13}$$

where ω_{10} is a scalar, $\boldsymbol{\omega}_{11}$ is a $B_L \times 1$ vector, and $\boldsymbol{\theta}^{LSTM} \equiv (\boldsymbol{\omega}, L, \mathbf{B})$ collectively denotes all parameters, including (i) the many weights in $\boldsymbol{\omega} \equiv \left(\left(\boldsymbol{\omega}_1^l, \boldsymbol{\omega}_2^l, \cdots, \boldsymbol{\omega}_9^l\right)_{l=1}^L, \omega_{10}, \boldsymbol{\omega}_{11}\right)$, (ii) the number of layers L, and (iii) the profile of the number of blocks in each layer, $\mathbf{B}$. We set $L = 3$ and $\mathbf{B} = (16, 8, 4)$, and find $\boldsymbol{\omega}$ that approximately maximizes the accuracy of prediction (to be explained in section 3.3 and Appendix B.2). In summary, LSTM sequentially processes the daily price data in a nonparametric Markov model, and uses the terminal state s^* as a latent score to detect cycles.

Method 9: Ensemble in Random Forests ("E-RF"). This method combines Methods 1–7 within random forests (RF), which is a class of nonparametric regressions. Let

$$g_{i,t}^m \equiv LHS_{i,t}^m - RHS_{i,t}^m \tag{14}$$

denote a "gap," the scalar difference between the left-hand side (LHS) and the right-hand side (RHS) of the inequality that defines each method $m = 1, 2, \cdots, M$, excluding the threshold parameter, $\boldsymbol{\theta}^m$. For example, inequality (2) defines Method 2. Hence, $g_{i,t}^2 = \left|mean_{d \in t}\left(\Delta p_{i,d}^+\right)\right| - \left|mean_{d \in t}\left(\Delta p_{i,d}^-\right)\right|$.[20] Let

$$\mathbf{g}_{i,t} \equiv \left(g_{i,t}^m\right)_{m=1}^M \tag{15}$$

[19] See Appendix B.2 for further details on this specification and computational implementation.

[20] All of Methods 1–7 except 4 are one-parameter models like this example. For Method 4, we define $g_{i,t}^4 \equiv \sum_{d \in t} \mathbb{I}\left\{\Delta p_{i,d} > \theta_1^{MBPI*}\right\}$, where θ_1^{MBPI*} is the accuracy-maximizing value of θ_1^{MBPI}.

denote their vector, where $M = 7$.[21] We construct a decision-trees classification algorithm that takes $\mathbf{g}_{i,t}$ as inputs and predicts $cycle_{i,t} = 1$ if and only if

$$h\left(\mathbf{g}_{i,t}; \boldsymbol{\omega}^{RF}, \boldsymbol{\kappa}^{RF}\right) \equiv \sum_{k=1}^{K} \omega_k^{RF} \mathbb{I}\left\{\mathbf{g}_{i,t} \in R_k\right\} \equiv \sum_{k=1}^{K} \omega_k^{RF} \phi\left(\mathbf{g}_{i,t}; \boldsymbol{\kappa}_k^{RF}\right) > 0, \tag{16}$$

where K is the number of adaptive basis functions, ω_k^{RF} is the weight of the k-th basis function, R_k is the k-th region in the M-dimensional space of $\mathbf{g}_{i,t}$, and $\boldsymbol{\kappa}_k^{RF}$ encodes both the choice of variables (elements of $\mathbf{g}_{i,t}$) and their threshold values that determine region R_k.[22] Because finding the truly optimal partitioning is computationally difficult, we use an RF algorithm to stochastically approximate it.[23] Thus, this method aggregates Methods 1–7 in a flexible manner that permits (i) multiple thresholds and (ii) interactions between $g_{i,t}^m$s. We denote its full set of parameters by $\boldsymbol{\theta}^{RF} \equiv \left(\boldsymbol{\omega}^{RF}, \boldsymbol{\kappa}^{RF}\right) \equiv \left(\left(\omega_k^{RF}\right)_{k=1}^{K}, \left(\boldsymbol{\kappa}_k^{RF}\right)_{k=1}^{K}\right)$.

Method 10: Ensemble in LSTM ("E-LSTM"). This method combines Methods 1–8 within an extended LSTM by incorporating $\mathbf{g}_{i,t}$ in (15) as additional variables in the laws of motion:

$$\mathbf{s}_d^l = \tanh\left(\mathbf{c}_d^l\right) \circ \Lambda\left(\boldsymbol{\omega}_1^l + \boldsymbol{\omega}_2^l \Delta p_d + \boldsymbol{\omega}_3^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{12}^l \mathbf{g}\right), \text{ and} \tag{17}$$

$$\begin{aligned} \mathbf{c}_d^l = {} & \tanh\left(\boldsymbol{\omega}_4^l + \boldsymbol{\omega}_5^l \Delta p_d + \boldsymbol{\omega}_6^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{13}^l \mathbf{g}\right) \circ \Lambda\left(\boldsymbol{\omega}_7^l + \boldsymbol{\omega}_8^l \Delta p_d + \boldsymbol{\omega}_9^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{14}^l \mathbf{g}\right) \\ & + \mathbf{c}_d^{l-1} \circ \left[1 - \Lambda\left(\boldsymbol{\omega}_7^l + \boldsymbol{\omega}_8^l \Delta p_d + \boldsymbol{\omega}_9^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{14}^l \mathbf{g}\right)\right], \end{aligned} \tag{18}$$

where $\left(\boldsymbol{\omega}_{12}^l, \boldsymbol{\omega}_{13}^l, \boldsymbol{\omega}_{14}^l\right)$ are $B_l \times M$ matrices of weight parameters for $\mathbf{g}_{i,t}$ (we suppress (i,t) subscript here). Other implementation details are the same as Method 8.

3.3 Optimization of Parameter Values ("Training")

Whereas the existing research typically calibrates (i.e., manually tunes) the threshold parameters, we optimize this process by choosing the parameter values that maximize accuracy, which we define as the percentage of correct predictions,

$$\%\text{ correct}\left(\boldsymbol{\theta}\right) \equiv \frac{\sum_{(i,t)} \mathbb{I}\left\{\widehat{cycle}_{i,t}\left(\boldsymbol{\theta}\right) = cycle_{i,t}\right\}}{\#\ all\ predictions} \times 100, \tag{19}$$

[21] Our computational implementation incorporates two additional variants of each of Methods 5–7 as well (explained in Appendix B.2). Hence, the eventual value of M is $7 + (2 \times 3) = 13$.

[22] See Murphy (2012, ch. 16) for an introduction to adaptive basis-function models including RF.

[23] See Appendix B.2 for further details.

where $\widehat{cycle}_{i,t}(\boldsymbol{\theta}) \in \{0,1\}$ is the algorithmic prediction for observation (i,t) at parameter value $\boldsymbol{\theta}$, and $cycle_{i,t} \in \{0,1\}$ is the manual classification label (data). We analogously define two types of prediction errors, "false negative" and "false positive," in Appendix B.3. Thus,

$$\boldsymbol{\theta}^* \equiv \arg\max_{\boldsymbol{\theta}} \quad \% \text{ correct}(\boldsymbol{\theta}) \tag{20}$$

characterizes the optimized (or "trained") model for each method.[24]

4 Results

Table 2 compares the accuracy of all methods as follows:

1. Randomly split each labeled dataset into an 80% "training" subsample and a 20% "testing" subsample.
2. Optimize the parameter values of each model in the 80% training subsample.
3. Assess its "out-of-sample" prediction accuracy in the 20% testing subsample.[25]
4. Repeat steps 1–3 101 times.[26]
5. Report the medians and standard deviations of the prediction results.

Panel I shows the results in WA, where clear-cut cycles are known to exist. Almost all methods achieve high accuracy near or above 90%.[27] The flexible, nonparametric models of Methods 8–10 do particularly well with above 99% accuracy. Meanwhile, CS lags behind with (a still respectable) 85%.[28]

Panel II reports the results in NSW. Cycle detection in NSW is not as easy as in WA, but most methods achieve near or above 80% accuracy. The nonparametric methods are top performers again (87%–90%), followed by MBPI and the spectral methods (81%–82%).

[24] See Appendix B.3 for further details and an alternative criterion.

[25] This cross-validation procedure is particularly important for the nonparametric models of Methods 8–10, which contain many parameters and could potentially "over-fit" the training subsample.

[26] An odd number of bootstrap sample-splits facilitates the selection of the medians in step 5.

[27] Byrne and de Roos (2019) set $\theta_1^{MBPI} = 6$ and $\theta_2^{MBPI} = 3.75$ in their study of WA, which are reasonably close to our accuracy-maximizing values, 5.05 and 5, respectively.

[28] Its parameter value, $\theta_{roots}^{CS} = 22.50$, suggests the model is trained to focus on shorter cycles with frequencies less than $90 \div \frac{22.5}{2} = 8$ days. Byrne and de Roos (2019) show both weekly and two-weekly cycles exist in WA. Thus, CS ends up ignoring the latter, longer cycles.

By contrast, CS (74%) and NMC (71%) make mostly degenerate predictions in which they classify virtually all observations as cycles.[29]

Panel III shows most methods fail in Germany, where cycles are more subtle and data are noisier (i.e., our RAs reach unanimous decisions less often).[30] E-LSTM is the only method that achieves accuracy near 80%, followed by E-RF (76%) and LSTM (75%). Somewhat surprisingly, CS (71%) outperforms all other parametric models; MBPI (65%) is the only existing method with non-degenerate predictions, presumably because it does not exclusively rely on asymmetry.[31] This profile of successful/failed methods suggests a sizable fraction of the German cycles does not conform to the Edgeworth-type asymmetry. Thus, a systematic investigation into Edgeworth cycles would require both nonparametric cycle detection and asymmetry-based classification.[32]

In summary, four findings emerge from Table 2. First, the four existing methods (Methods 1–4) work well in the clean data environments of Australia, but mostly fail in the noisier data from Germany. The spectral methods (Methods 5–6) show similar performances. Second, by contrast, CS (Method 7) underperforms most other methods when cycles are clear and regular but does relatively well in noisier cases. Third, LSTM (Method 8) is sufficiently flexible to capture both clear and noisy cycles: the most accurate stand-alone method. Fourth, the ensemble methods (Methods 9–10) effectively leverage the information content of Methods 1–8 and usually outperform all of them. The fact that E-RF performs so well is particularly interesting, because it simply aggregates the descriptive statistics from Methods 1–7 in a more flexible manner (i.e., permitting their interactions and multiple thresholds).

How Much Data Do We Need? The accuracy of cycle detection improves with the size of the training dataset, but the rate of improvement is different across methods. Figure 2 shows performances when we restrict the training dataset to only 0.1%, 1%, 5%, 10%, and so on of the available samples.

Methods 1–7 and 9 perform surprisingly well with only 0.1% of the data, which corresponds to 25, 10, and 36 observations in WA, NSW, and Germany, respectively. The labor

[29] The poor performance of NMC is surprising in three ways. First, it performed well in WA. Second, it is one of the most widely used methods in the literature. Third, other methods that similarly focus on asymmetry (PRNR and MIMD) do significantly better (78%–79%).

[30] Table 1 shows 71% of the NSW data is unanimously labeled as "cycling" by three RAs, whereas $9.4\% + 7.8\% = 17.2\%$ is labeled as such by only two or one RAs. In the German sample, only 39.6% is unanimously "cycling," whereas RAs disagree in $20.1\% + 17.6\% = 37.7\%$ of the data. Appendix C reports results based on "cleaner" subsamples that eliminate such disagreed observations.

[31] The parameter values of CS ($\theta^{CS}_{roots} = 24.50$) and MBPI ($\theta^{MBPI}_1 = 1.25$ and $\theta^{MBPI}_1 = 14$) suggest the German cycles are shorter, with lower amplitude and higher frequency, than the Australian ones.

[32] Appendix E summarizes our practical recommendations.

cost of human-generated labels is negligible for such small samples (US$3.51, US$2.84, and US$6.48, respectively).[33] Hence, these methods are extremely cost effective.

By contrast, Methods 8 and 10 contain many parameters and need more data. For example, E-LSTM's accuracy in NSW is below 50% when it uses only 10 observations (0.1% subsamples). Fortunately, their performances dramatically improve with 1% subsamples, and they start outperforming all other methods when 5% subsamples are used.[34] The "critical" sample size above which they perform the best is in the order of several hundred observations. The associated cost of manual labeling is only tens of RA hours, or a few hundred US dollars.[35] Thus, even though LSTM and E-LSTM require more data for a given accuracy level, their total costs are still surprisingly low, making them the highest-accuracy methods within a limited amount of resources.

[33] Based on the hourly wage of US$13.50 for undergraduate RA work at Yale University as of 2021.

[34] Strictly speaking, E-RF slightly outperforms E-LSTM in subsamples up to 40% in WA, although their mean differences are small relative to their standard deviations (see Tables 5 and 6 in Appendix C).

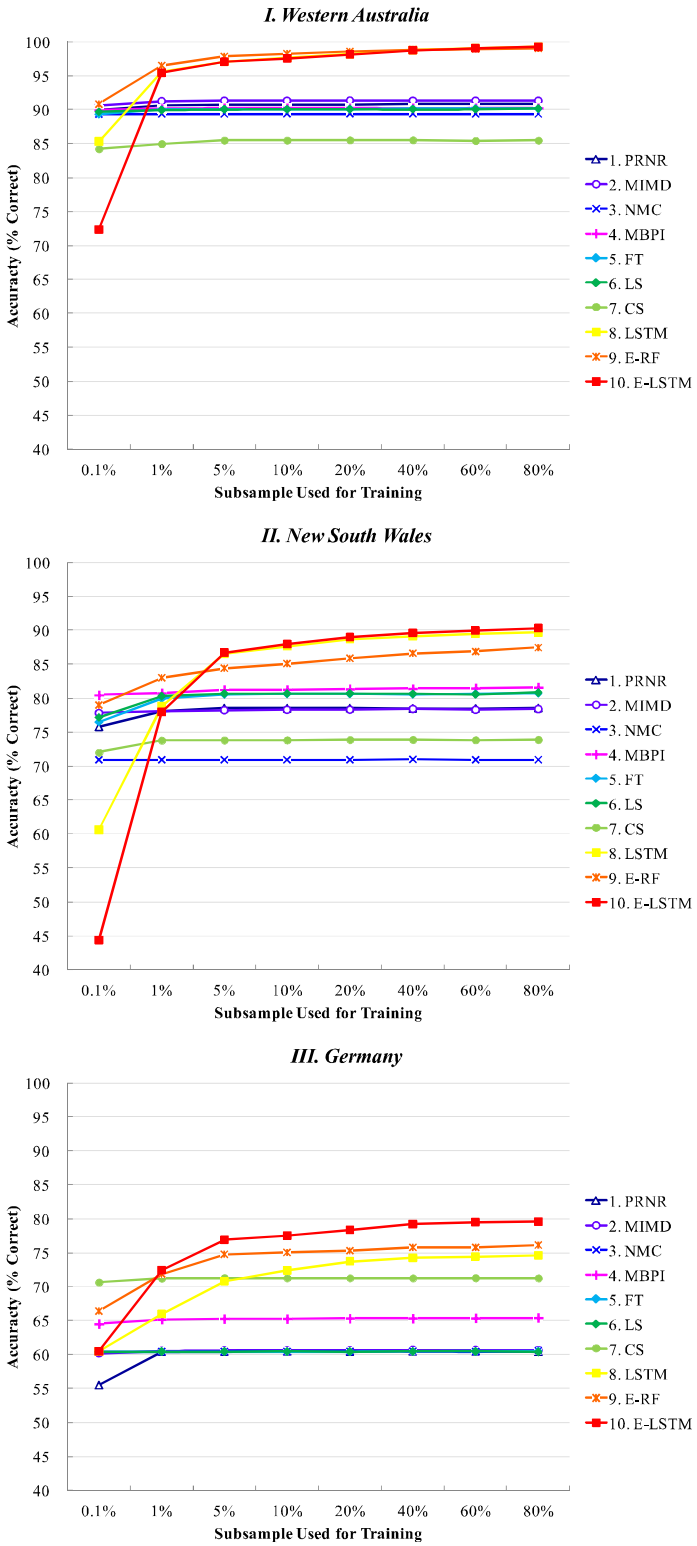
[35] Panel (B) of Table 5 in Appendix C reports the total costs of manual labeling for each dateset.

Table 2: Performance of Automatic Detection Methods

Method	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
I. Western Australia (# manually labeled observations: 24,569)										
Parameter 1	−1.16	6.13	−0.20	5.05	0.12	0.21	22.50	–	–	–
Parameter 2	–	–	–	5	–	–	–	–	–	–
Accuracy rank	5	4	9	6	8	7	10	1	3	1
% correct (median)	90.80	91.27	89.34	90.23	90.11	90.15	85.47	99.25	99.04	99.25
(Standard deviations)	(0.37)	(0.38)	(0.38)	(0.36)	(0.40)	(0.36)	(0.45)	(0.18)	(0.15)	(0.14)
of which cycling	55.27	55.70	57.08	60.74	58.24	57.92	56.41	60.62	60.97	60.34
of which not	35.53	35.57	32.25	29.49	31.87	32.23	29.06	38.62	38.07	38.91
% false negative	5.27	5.27	3.34	0.71	2.48	3.30	5.29	0.35	0.61	0.31
% false positive	3.93	3.46	7.33	9.06	7.41	6.55	9.24	0.41	0.35	0.45
II. New South Wales (# manually labeled observations: 9,693)										
Parameter 1	4.20	5.76	1.01	14.90	0.20	0.57	4.50	–	–	–
Parameter 2	–	–	–	2	–	–	–	–	–	–
Accuracy rank	7	8	10	4	6	5	9	2	3	1
% correct (median)	78.55	78.39	70.96	81.59	80.71	80.82	73.90	89.63	87.42	90.30
(Standard deviations)	(0.85)	(0.88)	(0.97)	(0.86)	(0.80)	(0.80)	(0.89)	(0.67)	(0.69)	(0.67)
of which cycling	67.04	65.09	70.96	64.62	66.53	66.43	70.40	67.20	67.10	65.60
of which not	11.50	13.31	0.00	16.97	14.18	14.39	3.51	22.43	20.32	24.70
% false negative	3.30	4.85	0.00	6.55	5.47	4.02	0.77	4.33	8.35	2.99
% false positive	18.15	16.76	29.04	11.86	13.82	15.16	25.32	6.03	4.23	6.70
III. Germany (# manually labeled observations: 35,685)										
Parameter 1	−3.48	0.30	−0.45	1.25	0.24	0.62	24.50	–	–	–
Parameter 2	–	–	–	14	–	–	–	–	–	–
Accuracy rank	9	6	7	5	8	10	4	3	2	1
% correct (median)	60.38	60.61	60.53	65.39	60.50	60.36	71.28	74.61	76.14	79.58
(Standard deviations)	(0.49)	(0.50)	(0.52)	(0.52)	(0.56)	(0.59)	(0.42)	(0.44)	(1.46)	(0.53)
of which cycling	0.00	1.25	0.07	14.77	0.00	0.00	25.88	23.46	23.96	29.96
of which not	60.38	59.37	60.46	50.62	60.50	60.36	45.40	51.16	52.18	49.63
% false negative	39.62	38.07	39.40	24.65	39.50	39.57	14.28	15.99	15.75	9.50
% false positive	0.00	1.32	0.07	9.96	0.00	0.07	14.45	9.40	8.11	10.91

Note: See section 3 for the definition of each method. Columns (8)–(10) do not report parameter values, because they contain too many parameters to be listed. We randomly split the sample into an 80% training subsample and a 20% testing subsample 101 times. In each split, the former subsample is used for setting parameter values, the medians of which are reported here. The accuracy statistics are also the medians from the 101 testing subsamples.

Figure 2: Gains from Additional Data



Note: The exact numbers underlying these plots are reported in Panel (A) of Table 5 in Appendix C.

5 Economic and Policy Implications

The suspicion that price cycles might be related to collusive business practices has led many researchers and governments to collect and scrutinize large data on fuel markets. Some papers find the presence of cycles is positively correlated with retail prices and markups, whereas others find the opposite relationships (see section 1). This section investigates how such findings depend on the definition of cycles.

Table 3 compares the retail-wholesale margins between "cycling" and "non-cycling" observations. Column (0) is based on our manual classification and serves as a "ground truth" benchmark. The mean margins in cycling and non-cycling observations in WA are A¢11.86 and A¢9.47, respectively. The mean difference is A¢2.39. The t test (based on Welch's t statistic) rejects the null hypothesis that the difference in means is zero at the 0.1% significance level. Hence, price cycles are positively correlated with margins in WA. The same analysis yields similar results in NSW. However, the pattern is reversed in Germany, where margins are *lower* in cycling station-quarters. Thus, in general, the presence of cycles (as recognized by human eyes) could be either positively or negatively correlated with margins, depending on regions/countries.

Columns (1)–(10) report the same analysis based on the 10 algorithmic methods. In WA, all methods reach the same conclusion that margins are higher in cycling observations. Broadly similar results emerge in NSW as well, even though one method fails (Method 3) and one reaches the opposite conclusion (Method 7). These discrepancies suggest whether researchers find a positive or negative cycle-margin relationship depends on the operational definition of cycles.

Our analysis of the German data highlights this point even more vividly. Both the manual classification and Methods 7–10 suggest significantly negative relationships between cycles and margins, but Methods 2–6 lead to *positive* mean differences. These positive differences are highly statistically significant in Methods 2–4. Some of them entail degenerate predictions (see section 4), but Method 4 features reasonable parameter values and achieves at least 65% accuracy. Hence, we cannot dismiss these discrepancies as purely random anomalies.

In summary, the choice of detection method could lead to qualitatively different results and dictate the policy implications of empirical research on Edgeworth cycles.

Table 3: Profit Margins by Cycle Status

Method	(0) Manual	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
I. Western Australia (# manually labeled observations: 24,569)											
Cycling											
# obs.	15,007	14,462	14,620	16,147	16,941	16,223	15,774	15,953	15,011	14,994	14,999
Mean	11.86	12.07	12.21	11.66	11.46	11.88	12.03	11.78	11.86	11.86	11.86
Std. dev.	4.01	3.80	3.74	3.98	4.13	3.87	3.85	4.04	4.01	4.01	4.01
Not cycling											
# obs.	9,562	10,107	9,949	8,422	7,628	8,346	8,795	8,616	9,558	9,575	9,570
Mean	9.47	9.30	9.05	9.52	9.73	9.08	8.94	9.35	9.47	9.47	9.47
Std. dev.	4.97	5.04	4.98	5.22	5.20	5.18	5.03	5.02	4.97	4.97	4.96
Difference											
Mean diff.	2.39	2.77	3.16	2.14	1.73	2.80	3.09	2.43	2.39	2.39	2.39
Welch's t	39.53	46.74	53.80	32.96	25.64	43.53	50.02	38.67	39.53	39.55	39.60
D. F.	17,247	17,771	17,314	13,648	12,134	13,263	14,608	14,723	17,236	17,282	17,295
p value	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
II. New South Wales (# manually labeled observations: 9,693)											
Cycling											
# obs.	6,878	8,324	8,038	9,693	7,303	7,704	7,994	9,253	7,052	6,961	7,183
Mean	12.03	11.73	12.35	11.66	12.48	11.76	11.81	11.58	12.19	12.07	12.13
Std. dev.	5.51	5.80	5.58	6.04	5.48	5.89	5.84	5.99	5.54	5.53	5.56
Not cycling											
# obs.	2,815	1,369	1,655	0	2,390	1,989	1,699	440	2,641	2,732	2,510
Mean	10.76	11.25	8.33	–	9.18	11.28	10.97	13.48	10.25	10.64	10.33
Std. dev.	7.10	7.31	7.01	–	6.92	6.56	6.85	6.79	7.01	7.08	7.08
Difference											
Mean diff.	1.27	0.48	4.02	–	3.30	0.48	0.84	−1.90	1.94	1.43	1.80
Welch's t	8.50	2.31	21.94	–	21.24	2.97	4.70	−5.76	12.80	9.48	11.55
D. F.	4,266	1,663	2,106	–	3,423	2,870	2,252	472	3,939	4,103	3,648
p value	$< .001$	.021	$< .001$	–	$< .001$	.003	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
III. Germany (# manually labeled observations: 35,685)											
Cycling											
# obs.	14,116	0	1,013	72	8,763	7	7	14,281	11,762	13,574	15,299
Mean	98.18	–	99.57	99.67	98.73	114.11	115.64	98.19	98.38	98.16	98.18
Std. dev.	3.57	–	6.96	3.26	3.84	32.10	31.40	3.60	3.60	3.59	3.51
Not cycling											
# obs.	21,569	35,685	34,672	35,613	26,922	35,678	35,678	21,404	23,923	22,111	20,386
Mean	98.65	98.46	98.43	98.46	98.38	98.46	98.46	98.65	98.50	98.65	98.68
Std. dev.	4.37	4.08	3.96	4.08	4.15	4.05	4.05	4.36	4.30	4.34	4.45
Difference											
Mean diff.	−0.47	–	1.14	1.21	0.35	15.65	17.18	−0.46	−0.12	−0.49	−0.50
Welch's t	−11.11	–	5.19	3.14	7.26	1.29	1.45	−10.86	−2.77	−11.55	−11.86
D. F.	33,984	–	1,031	71	15,941	6	6	34,110	27,415	32,697	35,595
p value	$< .001$	–	$< .001$	.002	$< .001$	.245	.197	$< .001$	.006	$< .001$	$< .001$

Note: Columns (1)–(10) use the median-accuracy version of each method in Table 2. The unit of measurement (of means and standard deviations) is the Australian cent in WA and NSW, and the euro cent in Germany, respectively. The p value indicates the probability that the difference in means is zero based on Welch's t statistic and the approximate degrees of freedom.

6 Discussions

This section discusses some of the findings and remaining questions.

Why Existing Methods Work in Australia But Fail in Germany. Most of the cycles in Australia exhibit asymmetry, whereas German cycles are not necessarily asymmetric (see Appendix D.1). Hence, asymmetry-based methods correctly identify cycles in Australia but not in Germany.

Why Margins And Cycles Correlate Positively in Australia But Negatively in Germany. In all datasets, the mean and the standard deviation of margins are positively correlated, because the only direction in which margins can move a lot is *upward* unless stations are willing to incur losses. We find volatility and cyclicality are correlated positively in Australia but negatively in Germany (see Appendix D.2). Therefore, the average level and cyclicality of margins are correlated positively in Australia but negatively in Germany.

How Can Cycles Be Less Volatile Than Non-Cycles? Cyclicality implies systematic—but not necessarily large—movements; not all large/frequent movements follow cycles. Many German observations exhibit high volatility without any discernible patterns.

Why Existing Methods Find "Positive Correlations." These methods' threshold rules tend to pick up high-mean, high-volatility cases as "cycles," because only sufficiently large movements can satisfy these conditions (see Appendix D.3). In Germany, however, volatility is a poor predictor of cyclicality (see above).

Manual Classification Focuses on "Cyclicality" but Not "Asymmetry." Maybe "Asymmetric Cycles" Do Predict Higher Markups Even in Germany? The answer is "no." The results are virtually the same when we focus on asymmetric cycles (see Appendix D.4).

Why Manual Classification Provides a Relevant Benchmark. If we had a perfect mathematical definition, no detection problem would arise. In the absence of such a formula, the existing research relied on rules of thumb that were ultimately validated by selective eyeballing. We made this process more systematic and transparent.

Moreover, even if we had a perfect definition, that would not stop consumers and politicians from eyeballing. In terms of political economy, it is their cognitions and reactions that

determine whether Edgeworth cycles become a public-policy issue. For these reasons, we believe a "ground truth" based on the consensus of reasonably well-educated student RAs provides a relevant benchmark.

7 Conclusion

This paper proposes a systematic approach to detect the widely studied Edgeworth-cycles phenomena, so that the growing amount of "big data" on fuel markets can be scrutinized in an accurate, scalable, and replicable manner. We formalize four existing methods, propose six new methods, and empirically assess their performance in the data from Australia and Germany.

Methodologically, five findings emerge. First, the difficulty of cycle detection varies by region/country. Second, all existing methods (Methods 1–4), which rely on the particular type of asymmetry that characterizes Edgeworth cycles, work well in WA and NSW but mostly fail in Germany, because not all of the German cycles conform to the Edgeworth-type asymmetry. These results suggest further investigation into Edgeworth cycles would benefit from the distinction between "asymmetry" and "cyclicality" in defining and detecting them. This paper offers a useful toolbox for these purposes.

Third, among our new methods, the performances of the spectral ones (Methods 5–6) are similar to those of the existing ones, whereas the splines-based approach (Method 7) is more robust to the noisier data in Germany. Fourth, the nonparametric models of Methods 8–10 detect cycles with the highest accuracy (approximately 99%, 90%, and 80% in WA, NSW, and Germany, respectively). Fifth, even though these complex models are generally known to require large training data, the requirement in the gas-price context is small: only a few hundred manually labeled observations (or tens of RA hours) are sufficient for even the most complex Method 10 to reach approximately optimal performances. Thus, our methods are effective as well as economical. Appendix E summarizes our practical recommendations.

Substantively, we show whether researchers discover a positive or negative statistical relationship between gas stations' profit margins and the existence of price cycles could critically depend on their choice of detection methods. Because the discovery of such "facts" lays the foundation for regulations and competition policy, these methodological considerations are directly policy relevant and consequential for the welfare performance of markets.

Appendix A Data and Manual-Classification Procedures

A.1 Data Sources and Preparation

Sources. *FuelWatch* and *FuelCheck* are legislated retail-fuel-price platforms operated by the state governments of WA and NSW, respectively. Their websites display real-time information on petrol prices, and the complete datasets can be downloaded.[36] The Market Transparency Unit for Fuels of the *Bundeskartellamt* publishes similar data for every German gas station in minute intervals.[37]

Sampling Frequencies. The raw data from WA contain daily prices for each station, which is the most granular level in this region, because its law mandates each station must commit to a fixed price level for 24 hours. By contrast, the stations in NSW and Germany can change prices at any point in time, which we aggregate into daily prices by taking either end-of-day values (NSW) or intra-day averages (Germany). Intra-day changes are relatively rare in NSW, and hence, end-of-day values are representative of the actual transaction prices. In Germany, many stations change prices multiple times during the day, so we sample 24 hourly prices and take their average for each station-day.

Wholesale Prices. The Australian Institute of Petroleum publishes average regional wholesale prices at https://www.aip.com.au. The Argus Media group's *OMR Oil Market Report* collects daily regional wholesale prices and offers the database on a commercial basis.

A.2 Manual-Classification Procedures

The manual labeling of the datasets proceeded in three stages. First, we labeled all station-quarters in the WA data with two RAs as a pilot project between July 2019 and June 2020. The first RA (a PhD student in economics) laid the ground work with approximately half of the WA data in close communication with one of the coauthors (Igami). The second RA (a senior undergraduate student majoring in economics) followed these examples to label the rest. Then, the first RA carefully double-checked all labels again to maintain consistency. As a result, each station-quarter (i, t) in WA has one label based on the consensus of the two RAs.

[36] Their URLs are https://www.fuelwatch.wa.gov.au and https://www.fuelcheck.nsw.gov.au.

[37] https://www.bundeskartellamt.de/EN/Economicsectors/MineralOil/MTU-Fuels/mtufuels_node.html

Second, the NSW dataset is smaller but contains more ambiguous cases. Hence, we took a more organized/computerized approach by building a cloud-based computational platform to streamline the labeling process. The same coauthor manually labeled a random sample of 100 station-quarters in December 2020, which is used for generating automated training sessions for three new undergraduate RAs (a senior and a junior majoring in economics, and a junior mathematics major). In the automated training sessions, each of the three RAs was asked to classify random subsamples of the labeled observations, and to repeat the labeling practice until their judgments agree with the coauthor's at least 80% of the time. Subsequently, each of them independently labeled the entire dataset in February–April 2021. Thus, each (i, t) in NSW carries three labels.

Third, the same team of three RAs proceeded to label a 5% random sample of the German dataset in April–June 2021. In turn, these labels served as a source of "training sample" for yet another team of three RAs (two juniors majoring in economics and a freshman in statistics and data science). They labeled an additional 5% random sample in June 2021. In total, 10% of the German data is triple-labeled.

In the computerized procedures for NSW and Germany, each RA is given a randomly selected observation for labeling at a time. Copying each other's answers would require keeping records of random sequences of thousands of observations. Hence, we believe the risk of collusion among RAs is low.

Appendix B Methodological Details

B.1 Other Existing Methods and Papers

The four existing methods (Methods 1–4) are based on the following studies:

1. Castanias and Johnson (1993) compare the lengths of positive and negative "runs" (defined as consecutive price increases and decreases, respectively) and find negative runs are longer than positive runs in the presence of cycles.

2. Eckert (2002) compares the magnitude of the mean increase and the mean decrease, and find the former is larger than the latter in the presence of cycles.[38]

3. Lewis (2009) proposes two indicators, both of which are commonly used in the subsequent literature. The first one is the median price change, which tends to be negative in

[38] Eckert (2003) uses this method as well. Clark and Houde (2014) propose its variant: the ratio of the median price increase to the median price decrease, with 2 as a threshold to define cyclical subsamples.

the presence of cycles.[39] The second measure is the number of price "jumps" (defined as price increases of at least 4 US cents per gallon in a single day or two consecutive days).[40]

4. Byrne and de Roos (2019) use a discretized version of Lewis's (2009) second indicator. They count the number of price jumps (defined as daily price increases of at least 6 Australian cents per liter), and define a gas station as "cycling" when many such jumps are observed (at least 15 times a year).

These papers are among the most cited in the literature, but the list is not exhaustive. Other influential papers use a variety of methods. Let us briefly discuss three of them. First, Noel (2007) proposes a Markov switching model with three unobserved states, two of which correspond to positive and negative runs, respectively, and the third corresponds to a non-cyclical regime.[41] Second, Deltas (2008) and many others regress retail price on wholesale price to describe asymmetric responses. Third, Foros and Steen (2013) regress price on days-of-week dummies to describe weekly cycles. These papers offer valuable insights, and their methods are suitable in their respective contexts. But they are not specifically designed for defining or detecting cycles, and hence, we do not include them in our framework.

B.2 Details of the New Methods

Fourier Transform (Method 5). The Fourier transform of a continuous function $g(x)$ is

$$G(f) \equiv \int_{-\infty}^{\infty} g(x)\, e^{-2\pi i f x} dx. \tag{21}$$

Let us define the Fourier transform operator $\mathcal{F}$ such that $\mathcal{F}\{g\} = G$, which is a linear operation. A sinusoidal signal (i.e., sine wave) with frequency f_0 has a Fourier transform

[39] Many studies use the median change, including Wills-Johnson and Bloch (2010), Doyle, Muehlegger, and Samphantharak (2010), Lewis and Noel (2011), Lewis (2012), Eckert and Eckert (2013), Zimmerman, Yun, and Taylor (2013), and Byrne (2019). As a threshold for discretization, Lewis (2012) uses -0.2 US cents per gallon, whereas Doyle et al. (2010) and Zimmerman et al. (2013) use -0.5 US cents per gallon.

[40] Lewis (2012) uses 5 US cents per gallon as an alternative threshold.

[41] Because these states are modeled as unobserved objects, using this approach as a definition is not straightforward. Zimmerman et al. (2013) propose another definition that shares the spirit of Markov switching regressions: (i) Compare the probabilities that a price increase (decrease) is observed after two consecutive price increases (decreases); and (ii) if the conditional probability of a third consecutive increase is smaller than that of a third decrease, take it as an indicator of cycles. We regard their approach as a variant of Castanias and Johnson's method. Finally, Noel (2018) defines the relenting and undercutting phases by consecutive days with cumulative increases and decreases of at least 3 Australian cents per liter, respectively, which is also close to Castanias and Johnson's (1993) idea.

consisting of a weighted sum of the Dirac delta functions at $\pm f_0$.[42] The practical implication of these properties is that any signal made up of a sum of sinusoidal components will have a Fourier transform consisting of a sum of delta functions that mark the frequencies of those sinusoids. Thus, the Fourier transform directly measures additive periodic content in a continuous function. The power spectral density (PSD, or the power spectrum) of a function,

$$P_g \equiv \left|\mathcal{F}\left\{g\right\}\right|^2, \tag{22}$$

is a positive, real-valued function of frequency f, and provides a convenient way to quantify the contribution of each frequency f to the signal $g\left(x\right)$.

When a continuous time series is sampled at regular time intervals with spacing Δx, as is the case in our data, one can use the discrete version of (21):

$$G_{obs}\left(f\right) = \sum_{n=-\infty}^{\infty} g\left(n\Delta x\right) e^{-2\pi i f n \Delta x}. \tag{23}$$

Acknowledging the finite sample size N and focusing on the relevant frequency range $0 \leq f \leq \frac{1}{\Delta x}$, one can define N evenly spaced frequencies with $\Delta f = \frac{1}{N\Delta x}$ covering this range. Let $g_n \equiv g\left(n\Delta x\right)$ and $G_k \equiv G_{obs}\left(k\Delta f\right)$. Then, the sample analog of (21) is

$$G_k = \sum_{n=0}^{N} g_n e^{-2\pi i k n/N}. \tag{24}$$

One can construct the sample analog of the Fourier power spectrum (22) as (5) in the main text. This is the "classical" or "Schuster" periodogram.[43]

A potential drawback of the threshold rule in (6) is that it exclusively focuses on the highest point and ignores the rest. As an alternative rule, we can compare the highest point with the heights of other, less powerful frequencies. One way to capture relative heights of multiple frequencies is to measure the "concentration" of power in a limited number of frequencies. We use the Herfindahl-Hirschman Index (HHI) for an additional check for

[42] The Dirac delta function is $\delta\left(f\right) \equiv \int_{-\infty}^{\infty} e^{-2\pi i f x} dx$, and hence, we can write $\mathcal{F}\left\{e^{2\pi f_0 x}\right\} = \delta\left(f - f_0\right)$. The linearity of $\mathcal{F}$ and Euler's formula for the complex exponential ($e^{ix} = \cos x + i \sin x$) lead to the following identities: $\mathcal{F}\left\{\cos\left(2\pi f_0 x\right)\right\} = \frac{1}{2}\left[\delta\left(f - f_0\right) + \delta\left(f + f_0\right)\right]$ and $\mathcal{F}\left\{\sin\left(2\pi f_0 x\right)\right\} = \frac{1}{2i}\left[\delta\left(f - f_0\right) + \delta\left(f + f_0\right)\right]$. See VanderPlas (2018) for further details.

[43] See Press et al. (1992, section 12.2) for computational implementation.

"significant" cycles:

$$HHI_{i,t} \equiv \sum_f \left(\frac{P_{i,t}(f)}{\sum_f P_{i,t}(f)} \right)^2 > \theta_{hhi}^{FT}, \tag{25}$$

where $\theta_{hhi}^{FT} \in (0,1]$ is a scalar threshold parameter.[44] A high value of $HHI_{i,t}$ indicates strong periodicity at certain frequencies relative to other, weaker frequencies.

Lomb-Scargle Periodogram (Method 6). Even though the classical periodogram in (5) appears different from (7), (5) can be rewritten as

$$P(f) = \frac{1}{N} \left[\left(\sum_n g_n \cos(2\pi f x_n) \right)^2 + \left(\sum_n g_n \sin(2\pi f x_n) \right)^2 \right].$$

Thus, the only major difference between (5) and (7) is the denominators in (7).

Statistically, one can interpret the Lomb-Scargle periodogram as a collection of least-squares regressions in which one fits a sinusoidal model at each frequency f:

$$\hat{g}(x; f) = A_f \sin\left(2\pi f \left(x - \phi_f\right)\right), \tag{26}$$

where amplitude A_f and phase ϕ_f are the parameters to be estimated by minimizing the sum of squared residuals:

$$SSR^{LS}(f) \equiv \sum_n \left(g_n - \hat{g}(x_n; f)\right)^2. \tag{27}$$

Scargle (1982) shows the following periodogram is identical to (7):

$$\tilde{P}^{LS}(f) = \frac{1}{2} \left[SSR_0^{LS} - SSR^{LS}(f) \right],$$

where SSR_0^{LS} is the sum of squared residuals from the restricted model in which the only regressor is a constant term. The idea is that the frequencies with good fit will exhibit high $\tilde{P}^{LS}(f)$.

[44] The HHI is a summary statistic that is typically used to measure the degree of market-share concentration in oligopolistic industries. A high value of the HHI indicates the market is close to monopoly.

The HHI variant of the LS method is

$$HHI_{i,t}^{LS} \equiv \sum_{f} \left(\frac{P_{i,t}^{LS}(f)}{\sum_f P_{i,t}^{LS}(f)} \right)^2 > \theta_{hhi}^{LS}. \tag{28}$$

Cubic Splines (Method 7). A spline is a piecewise polynomial function:

$$S_K(x) = \sum_{j=0}^{P} \beta_j x^j + \sum_{k=1}^{N} \beta_{P+k} (x - \tau_k)^P \mathbb{I}\{x \geq \tau_k\}, \tag{29}$$

where $K = 1 + P + N$ is the number of coefficients, P is the order of the polynomial (not to be confused with the periodogram in Methods 5–6 or our notation for the price, p), and the support for x is covered by $N + 1$ ordered subintervals that are joined by N knots $(\tau_1 < \tau_2 < \cdots < \tau_N)$.[45] It is a special case of a sieve/series approximation that constitutes a class of nonparametric regression methods.[46] We use splines as an interpolator to smooth the discrete (daily) time series and facilitate further calculations. Specifically, we use a cubic Hermite interpolator, which is a spline where each piece is a third-degree polynomial of Hermite form (i.e., $P = 3$, $N = 88$, and βs are prespecified).[47]

In addition to the indicator of frequent oscillations in (10), we propose a measure that captures amplitude as well. We subtract the lowest daily price in (i, t) from all of its daily prices, $\underline{p}_{i,d} \equiv p_{i,d} - min_{d \in t}(p_{i,d})$, fit CS to $\left(\underline{p}_{i,d}\right)_{d \in t}$, and calculate its integral over $d \in [1, 90]$. We set $cycle_{i,t} = 1$ if and only if

$$\int_1^{90} \underline{CS}_{i,t}(d) > \theta_{int}^{CS}, \tag{30}$$

where $\underline{CS}_{i,t}(d)$ is the fitted value of $\underline{p}_{i,d}$ at time d. Because this definite integral equals the area between the price series and its lowest level within (i, t), this condition captures cycles with large amplitude and sustained high prices.

[45] This N should not be confused with our notation for sample size in the discrete Fourier transform.

[46] Any continuous function can be uniformly well approximated by a polynomial of sufficiently high order, and the rate of approximation is $o\left(K^{-2}\right)$. Other series models include trigonometric polynomials, wavelets, orthogonal wavelets, B-splines, and artificial neural networks. See Hansen (2020, ch. 20) for an introduction and Chen (2007) for a review.

[47] On the unit interval $d \in (0, 1)$, given a starting point p_0 at $d = 0$, an ending point p_1 at $d = 1$, and slopes m_0 and m_1, this polynomial is

$$p(d) = \left(2d^3 - 3d^2 + 1\right) p_0 + \left(d^3 - 2d^2 + d\right) m_0 + \left(-2d^3 + 3d^2\right) p_1 + \left(d^3 - d^2\right) m_1.$$

This form ensures the observed values (p_0, p_1) and their slopes (m_0, m_1) are fitted exactly. It has become a default specification of CS in SciPy, a set of commonly used Python libraries for scientific computing.

We also construct a discrete (raw data) analog of the splines-integral measure as follows:

$$\sum_{d=1}^{90} |\bar{p}_{i,d}| > \theta_{abs}^{CS}, \tag{31}$$

where $\bar{p}_{i,d}$ is the demeaned price. The information content of this statistic is similar to the previous one, but its calaculation is simpler.

Long Short-Term Memory (Method 8). Compared with Greff et al.'s (2017) "vanilla" setup, we make two simplifications. First, our law of motion for $\mathbf{c}_d^l$ (12) uses the same set of parameters $\left(\boldsymbol{\omega}_7^l, \boldsymbol{\omega}_8^l, \boldsymbol{\omega}_9^l\right)$ twice. This simplification corresponds to their "Coupled Input and Forget Gate" variant due to Cho et al. (2014), which is also referred to as Gated Recurrent Units (GRUs) in the literature. Second, we do not include $\mathbf{c}_d^l$ or $\mathbf{c}_{d-1}^l$ inside Λ in (11) or inside tanh and Λ in (12). This omission corresponds to their "No Peepholes" variant. Greff et al. (2017) show these simplifications reduce the number of parameters without compromising predictive accuracy.

We implement LSTM in TensorFlow-GPU 2.6 (tf.keras.models.Sequential). Our choice of network architecture and activation functions—which constitute the specification of effective functional forms—are as explained in the main text. The total number of weight parameters is 2,165. We set other tuning parameters and the details of numerical optimization as follows: (i) the dropout rate is 0.5, (ii) the optimizer is tf.keras.optimizer.RMSprop with the learning rate of 0.0005, (iii) the number of epochs is 100, and (iv) the batch size is 30.

Ensemble in Random Forests (Method 9). The relationship between "decision trees" and "random forests" is as follows, according to Murphy (2012, ch. 16). Because finding the truly optimal partitioning in a decision-trees model is computationally infeasible, some greedy, iterative procedures are used in the estimation/tuning of the parameters $\left(\boldsymbol{\omega}^{RF}, \boldsymbol{\kappa}^{RF}\right)$. However, the hierarchical nature of this process leads to unstable predictions. Averaging over multiple estimates from bootstrapped subsamples ("bootstrap aggregating" or "bagging") is a commonly used technique to reduce this variance. A further improvement is possible by randomly choosing a subset of input variables, in addition to "bagging." This technique is called "random forests" (Breiman 2001a) and is known to perform well in many different contexts (e.g., Caruana and Niculescu-Mizil 2006).

We implement E-RF in scikit-learn 0.24.2 (sklearn.ensemble.RandomForestClassifier), with default options for all settings.

Ensemble in Long Short-Term Memory (Method 10). Our E-LSTM implementation details are the same as in the basic LSTM (Method 8). The only difference is that the total number of weight parameters is larger at 2,933 to incorporate the additional input variables from Methods 1–7.

B.3 Parameter Optimization

We define two types of prediction errors as follows:

$$\%\text{ false negative}\left(\boldsymbol{\theta}\right) \equiv \frac{\sum_{(i,t)} \mathbb{I}\left\{\widehat{cycle}_{i,t}\left(\boldsymbol{\theta}\right)=0, cycle_{i,t}=1\right\}}{\#\ all\ predictions} \times 100, \text{ and} \tag{32}$$

$$\%\text{ false positive}\left(\boldsymbol{\theta}\right) \equiv \frac{\sum_{(i,t)} \mathbb{I}\left\{\widehat{cycle}_{i,t}\left(\boldsymbol{\theta}\right)=1, cycle_{i,t}=0\right\}}{\#\ all\ predictions} \times 100. \tag{33}$$

They correspond to type II errors and type I errors in statistics, respectively.

We occasionally encounter cases in which a range of parameter values attain the same (maximum) accuracy. In such cases, we report the median of all $\boldsymbol{\theta}^*$ values that we find in our grid search. These cases typically involve "degenerate" predictions in which $\widehat{cycle}_{i,t}\left(\boldsymbol{\theta}\right)=1$ or $\widehat{cycle}_{i,t}\left(\boldsymbol{\theta}\right)=0$ for all (i,t), and hence are mostly irrelevant for the purpose of finding well-performing $\boldsymbol{\theta}$s.

Maximum-likelihood estimation (MLE) offers an alternative to accuracy maximization that avoids multiple optima, but we do not pursue MLE in the main text, because the likelihood-maximizing parameter values lead to too many degenerate predictions in our data context.

The MLE version of our parameter tuning (or "training") procedures uses a binary logit model that incorporates each of the definitions in section 3. The probability that human eyes and brains detect cycles in the vector of daily price data for station i in quarter t is

$$\Pr\left(cycle_{i,t}=1|\mathbf{p}_{i,t};\boldsymbol{\theta}_0\right)=\Lambda\left(g\left(\boldsymbol{\theta}_0\right)\right), \tag{34}$$

and $\Pr\left(cycle_{i,t}=0|\mathbf{p}_{i,t};\boldsymbol{\theta}_0\right)=1-\Lambda\left(g\left(\boldsymbol{\theta}_0\right)\right)$, where $\mathbf{p}_{i,t}\equiv\left[p_{i,d}\right]_{d=1}^{D(t)}$ is a $D\left(t\right)$ vector of daily prices, $D\left(t\right)$ is the number of days in quarter t, $\boldsymbol{\theta}_0$ is the true parameter value (either a scalar or a vector, depending on the method) in the data-generating process (DGP), $g\left(\boldsymbol{\theta}_0\right)\equiv LHS-RHS$ of the definitional inequality of each method (i.e., inequalities 1,

2, 3, ...), and Λ is the cumulative density function (CDF) of the logistic distribution:

$$\Lambda(g) \equiv \frac{\exp(g)}{1+\exp(g)}. \tag{35}$$

The probability density function (PDF) can be written as

$$l(cycle_{i,t}|\mathbf{p}_{i,t};\boldsymbol{\theta}_0) = \Lambda(g(\boldsymbol{\theta}_0))^{cycle_{i,t}} \times [1-\Lambda(g(\boldsymbol{\theta}_0))]^{1-cycle_{i,t}}. \tag{36}$$

Taking logs of both sides and replacing $\boldsymbol{\theta}_0$ by its hypothetical value $\boldsymbol{\theta}$, we obtain the log likelihood for observation (i,t):

$$\log l(cycle_{i,t}|\mathbf{p}_{i,t};\boldsymbol{\theta}) = cycle_{i,t} \times \log\Lambda(g(\boldsymbol{\theta})) + (1-cycle_{i,t}) \times \log[1-\Lambda(g(\boldsymbol{\theta}))]. \tag{37}$$

Under the assumption that $\{cycle_{i,t}, \mathbf{p}_{i,t}\}$ is i.i.d. across i and t,[48] the log likelihood of the manually labeled sample is the sum of (37) over i and t. Therefore, the maximum-likelihood estimator of $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} \sum_i \sum_t \log l(cycle_{i,t}|\mathbf{p}_{i,t};\boldsymbol{\theta}). \tag{38}$$

Appendix C Additional Results

Variants of FT, LS, and CS. Table 4 reports the performances of the variants of Methods 5, 6, and 7. In Methods 5 and 6, the "max" and "HHI" variants are as explained in section 3.2 and Appendix B.2. The "peak" variant is similar to the "max" one except that we additionally use a peak-detection algorithm to ensure we are measuring the height of the highest (and well-behaved) peak in the power spectrum and not some accidental maximum due to noisy data. In Method 7, the "roots" variant is the baseline version in section 3.2. Its "integral" and "absolute value" variants are explained in Appendix B.2.

Data Requirement and Marginal Cost of Accuracty. Table 5 reports the means of accuracy (% correct) across 101 bootstrap sample splits that are underlying the visual summaries in Figure 2 in section 5. Table 6 shows the standard deviations of accuracy are

[48] Note $\boldsymbol{p}_{i,t}$ is a vector of prices for 90 consecutive days. To the extent that these daily prices are serially correlated, this assumption may not hold for $\boldsymbol{p}_{i,t}$ in practice. Nevertheless, our manual labeling of each $\boldsymbol{p}_{i,t}$ (i.e., the DGP for $cycle_{i,t}$) does proceed with a random-sampling procedure that is i.i.d. across i and t. Hence, the i.i.d. assumption is an accurate description of the DGP for $cycle_{i,t}$, our main prediction target.

usually less than 1 percentage point when more than 1% of the sample is used for training.

Figure 3 and Panel (C) of Table 5 show the "marginal costs of accuracy" (i.e., the amount of RA work required for an extra percentage-point increase in accuracy). The marginal cost is initially low with only a few cents, but rapidly increases as we approach the maximum possible accuracy levels. Because the difficulty of accurate classification in a new dataset is unknown *a priori*, one cannot set realistic targets without some preliminary analysis. Nevertheless, our findings in section 4 are encouraging in that only a few hundred labeled observations are necessary to reach approximately optimal accuracy levels.

Table 4: Performance of Automatic Detection Methods (Other Variants)

Method	(5) FT_{max}	(5′) FT_{peak}	(5″) $FT2_{hhi}$	(6) LS_{max}	(6′) LS_{peak}	(6″) LS_{hhi}	(7) CS_{roots}	(7′) CS_{int}	(7″) CS_{abs}
I. Western Australia (# manually labeled observations: 24,569)									
Parameter 1	0.12	0.14	0.04	0.21	0.23	0.44	22.50	551.47	246.08
Parameter 2	–	–	–	–	–	–	–	–	–
% correct (median)	90.11	88.40	87.61	90.15	89.66	81.83	85.47	83.42	85.14
(Standard deviations)	(0.40)	(0.45)	(0.39)	(0.36)	(0.43)	(0.54)	(0.45)	(0.54)	(0.42)
of which cycling	58.24	57.31	59.12	57.92	57.10	54.13	56.41	55.92	57.14
of which not	31.87	31.09	28.49	32.23	32.56	27.70	29.06	27.49	28.00
% false negative	2.48	4.05	1.91	3.30	4.50	6.15	5.29	4.82	3.32
% false positive	7.41	7.5	10.48	6.55	5.84	12.03	9.24	11.76	11.54
II. New South Wales (# manually labeled observations: 9,693)									
Parameter 1	0.20	0.27	0.21	0.57	0.81	29.21	4.50	783.11	459.83
Parameter 2	–	–	–	–	–	–	–	–	–
% correct (median)	80.71	81.85	81.23	80.82	82.21	81.38	73.90	75.45	79.63
(Standard deviations)	(0.80)	(0.70)	(0.83)	(0.80)	(0.81)	(0.84)	(0.89)	(0.79)	(0.87)
of which cycling	66.53	66.99	64.72	66.43	66.89	67.30	70.40	68.13	67.25
of which not	14.18	14.85	16.50	14.39	15.32	14.08	3.51	7.32	12.38
% false negative	5.47	4.54	6.19	4.02	4.07	4.54	0.77	3.20	3.56
% false positive	13.82	13.62	12.58	15.16	13.72	14.08	25.32	21.35	16.81
III. Germany (# manually labeled observations: 35,685)									
Parameter 1	0.24	0.90	0.67	0.62	1.93	42.96	24.50	994.19	4,623
Parameter 2	–	–	–	–	–	–	–	–	–
% correct (median)	60.50	60.57	60.35	60.36	60.50	60.52	71.28	60.29	60.49
(Standard deviations)	(0.56)	(0.50)	(0.53)	(0.59)	(0.57)	(0.53)	(0.42)	(0.48)	(0.48)
of which cycling	0.00	0.00	0.00	0.00	0.00	24.30	25.88	0.00	0.00
of which not	60.50	60.57	60.35	60.36	60.50	47.00	45.40	60.29	60.49
% false negative	39.50	39.41	39.65	39.57	39.50	15.26	14.28	39.51	0.00
% false positive	0.00	0.01	0.00	0.07	0.00	13.43	14.45	0.20	39.51

Note: See the text of Appendix C for the definition of each method.

Table 5: Benefits and Costs of Additional Data

Subsample used for "training"	(1) 0.1%	(2) 1%	(3) 5%	(4) 10%	(5) 20%	(6) 40%	(7) 60%	(8) 80%
(A) Median Accuracy (% correct)								
I. Western Australia (# manually labeled observations: 24,569)								
1. PRNR	89.88	90.64	90.70	90.73	90.72	90.80	90.80	90.80
2. MIMD	90.55	91.21	91.29	91.28	91.31	91.29	91.29	91.27
3. NMC	89.33	89.35	89.34	89.36	89.36	89.36	89.35	89.34
4. MBPI	89.95	90.15	90.24	90.27	90.26	90.25	90.26	90.23
5. FT	89.29	89.98	90.06	90.06	90.08	90.12	90.15	90.11
6. LS	89.67	89.92	90.01	90.01	90.03	90.06	90.06	90.15
7. CS	84.22	84.96	85.48	85.48	85.51	85.52	85.43	85.47
8. LSTM	85.32	95.54	97.11	97.66	98.20	98.78	99.07	99.25
9. E-RF	90.78	96.53	97.86	98.27	98.59	98.87	98.96	99.04
10. E-LSTM	72.36	95.46	97.09	97.59	98.15	98.76	99.06	99.25
II. New South Wales (# manually labeled observations: 9,693)								
1. PRNR	75.78	78.12	78.51	78.50	78.54	78.49	78.49	78.55
2. MIMD	77.81	78.13	78.24	78.29	78.31	78.42	78.29	78.39
3. NMC	70.95	70.95	70.95	70.93	70.94	70.99	70.94	70.96
4. MBPI	80.47	80.74	81.25	81.27	81.35	81.40	81.48	81.59
5. FT	76.48	79.93	80.56	80.62	80.64	80.57	80.63	80.71
6. LS	77.17	80.29	80.66	80.67	80.70	80.64	80.58	80.82
7. CS	72.02	73.80	73.83	73.84	73.89	73.90	73.85	73.90
8. LSTM	60.60	78.84	86.52	87.64	88.68	89.13	89.45	89.63
9. E-RF	78.99	83.00	84.40	85.09	85.85	86.59	86.85	87.42
10. E-LSTM	44.39	77.96	86.72	87.98	89.01	89.60	89.94	90.30
III. Germany (# manually labeled observations: 35,685)								
1. PRNR	55.48	60.43	60.42	60.46	60.42	60.45	60.47	60.38
2. MIMD	60.19	60.51	60.58	60.58	60.60	60.66	60.63	60.61
3. NMC	60.42	60.43	60.44	60.44	60.46	60.47	60.43	60.53
4. MBPI	64.49	65.16	65.25	65.26	65.32	65.32	65.35	65.39
5. FT	60.42	60.42	60.42	60.43	60.40	60.46	60.45	60.50
6. LS	60.42	60.42	60.43	60.44	60.42	60.44	60.46	60.36
7. CS	70.66	71.26	71.31	71.29	71.29	71.28	71.31	71.28
8. LSTM	60.45	65.97	70.83	72.40	73.74	74.27	74.43	74.61
9. E-RF	66.38	71.86	74.78	75.04	75.30	75.79	75.79	76.14
10. E-LSTM	60.45	72.43	76.91	77.50	78.38	79.21	79.51	79.58
(B) Total Costs of Manual Labeling (US$)								
I. Western Australia	3.51	35.1	176	351	702	1,404	2,106	2,808
II. New South Wales	2.84	28.4	142	284	567	1,134	1,701	2,268
III. Germany	6.48	64.8	324	648	1,296	2,592	3,888	5,184
(C) E-LSTM's Marginal Costs of Accuracy (US$ per correct % point)								
I. Western Australia	0.05	1.37	86.13	351	627	1,151	2,340	3,695
II. New South Wales	0.06	0.76	12.95	113	275	961	1,668	1,575
III. Germany	0.11	4.87	57.86	549	736	1,561	4,320	18,514

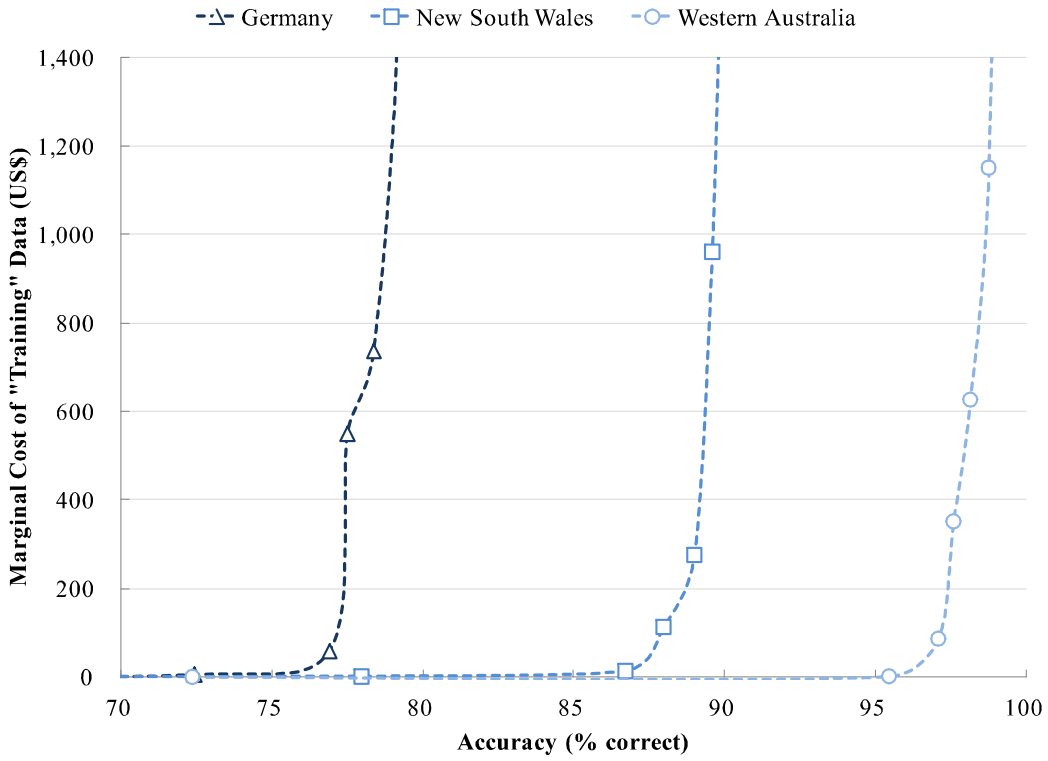
Note: The numbers in panel (A) indicate accuracy in the 20% testing subsample. Some or all of the remaining 80% subsample are used as a training subsample to optimize the parameters of each model. We randomly split the sample 101 times, tune the parameters as many times, and report their median performances. The dollar costs in Panels (B) and (C) are based on the total RA hours to manually classify cycles and the hourly wage of $13.50 (see section 2).

Table 6: Standard Deviations of Accuracy under Different Sample Sizes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Subsample used for "training"	0.1%	1%	5%	10%	20%	40%	60%	80%
Standard Deviation of: Accuracy (% correct)								
I. Western Australia (# manually labeled observations: 24,569)								
1. PRNR	1.94	0.64	0.33	0.11	0.12	0.16	0.22	0.37
2. MIMD	2.09	0.6	0.19	0.10	0.11	0.15	0.20	0.38
3. NMC	3.22	0.39	0.11	0.08	0.09	0.15	0.23	0.38
4. MBPI	2.34	0.46	0.11	0.09	0.11	0.14	0.22	0.36
5. FT	1.80	0.50	0.25	0.16	0.14	0.16	0.23	0.40
6. LS	1.39	0.38	0.14	0.15	0.12	0.16	0.24	0.36
7. CS	2.15	0.59	0.16	0.09	0.12	0.20	0.26	0.45
8. LSTM	23.81	1.15	0.20	0.21	0.20	0.21	6.01	0.18
9. E-RF	1.65	0.44	0.17	0.15	0.11	0.09	0.09	0.15
10. E-LSTM	23.22	1.32	0.21	0.19	5.91	0.20	0.16	0.14
II. New South Wales (# manually labeled observations: 9,693)								
1. PRNR	4.94	1.03	0.62	0.53	0.28	0.32	0.49	0.85
2. MIMD	1.90	0.98	0.47	0.28	0.30	0.36	0.52	0.88
3. NMC	12.86	0.04	0.10	0.16	0.21	0.36	0.56	0.97
4. MBPI	2.44	1.51	0.31	0.31	0.25	0.37	0.55	0.86
5. FT	6.57	1.65	0.54	0.36	0.30	0.40	0.48	0.80
6. LS	4.46	1.10	0.51	0.34	0.27	0.31	0.48	0.80
7. CS	14.34	1.01	0.43	0.17	0.27	0.41	0.61	0.89
8. LSTM	20.66	2.57	0.74	0.46	0.29	0.33	0.41	0.67
9. E-RF	6.34	0.94	0.42	0.36	0.34	0.37	0.46	0.69
10. E-LSTM	19.25	5.21	0.89	0.45	0.34	0.34	0.42	0.67
III. Germany (# manually labeled observations: 35,685)								
1. PRNR	4.35	2.06	0.11	0.08	0.12	0.21	0.31	0.49
2. MIMD	5.30	0.75	0.20	0.28	0.14	0.19	0.32	0.50
3. NMC	5.67	0.38	0.10	0.09	0.14	0.19	0.33	0.52
4. MBPI	1.64	0.84	0.27	0.19	0.16	0.22	0.28	0.52
5. FT	5.71	0.23	0.11	0.09	0.14	0.20	0.30	0.56
6. LS	6.97	0.15	0.07	0.10	0.15	0.25	0.33	0.59
7. CS	2.98	0.70	0.20	0.10	0.13	0.17	0.29	0.42
8. LSTM	0.68	1.47	1.05	1.08	0.67	0.40	0.38	0.44
9. E-RF	2.99	2.29	2.26	1.33	1.47	1.29	1.47	1.46
10. E-LSTM	0.80	2.08	0.62	0.47	0.63	0.51	0.41	0.53

Note: The numbers indicate the standard deviations of accuracy across the 101 bootstrap sample-splits.

Figure 3: Marginal Costs of Accuracy (E-LSTM)



Note: These marginal-cost curves are based on the numbers reported in Table 5 (shown with markers) and a splines-based interpolation (dashed lines).

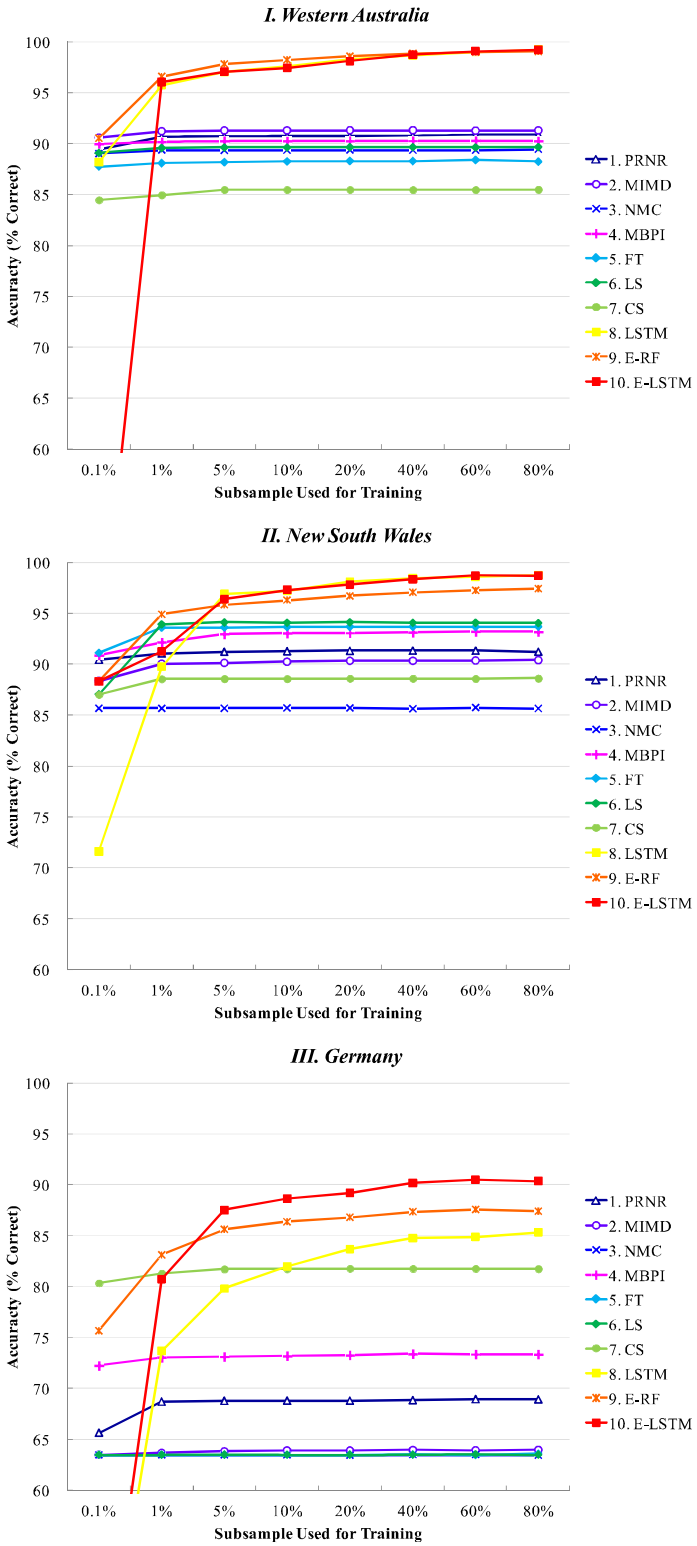
Using Only "Cleaner" Subsamples. Table 7 and Figure 4 report alternative results based only on subsamples that are either unanimously labeled as "cycling" by all RAs or not labeled as "cycling" by any RA, thereby ignoring ambiguous observations. Accuracy is higher overall, but relative performance rankings remain similar to the baseline results.

Table 7: Performance of Automatic Detection Methods in "Cleaner" Datasets

Method	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
I. Western Australia (# manually labeled observations: 24,569)										
Parameter 1	−1.16	5.14	−0.20	4.85	0.14	0.23	22.50	–	–	–
Parameter 2	–	–	–	5	–	–	–	–	–	–
Accuracy rank	5	4	8	6	9	7	10	1	3	2
% correct (median)	90.86	91.29	89.44	90.25	88.22	89.66	85.49	99.25	99.06	99.21
(Standard deviations)	(0.29)	(0.33)	(0.43)	(0.37)	(0.39)	(0.42)	(0.51)	(0.16)	(0.14)	(0.12)
of which cycling	55.45	56.98	58.26	59.69	57.22	56.61	55.80	60.09	60.44	60.68
of which not	35.41	34.31	31.18	30.57	30.99	33.05	29.69	39.15	38.62	38.52
% false negative	5.66	3.93	3.01	0.53	3.93	3.64	5.33	0.53	0.55	0.49
% false positive	3.48	4.78	7.55	9.22	7.86	6.70	9.18	0.22	0.39	0.31
II. New South Wales (# manually labeled observations: 8,028)										
Parameter 1	5.24	2.73	1.01	8.9	0.22	0.57	4.50	–	–	–
Parameter 2	–	–	–	2	–	–	–	–	–	–
Accuracy rank	7	8	10	6	5	4	9	1	3	2
% correct (median)	91.22	90.39	85.65	93.17	93.68	94.03	88.61	98.75	97.44	98.70
(Standard deviations)	(0.66)	(0.63)	(0.75)	(0.49)	(0.62)	(0.48)	(0.66)	(0.24)	(0.37)	(0.37)
of which cycling	84.18	84.39	85.65	84.18	85.02	83.53	84.63	84.56	85.06	85.18
of which not	7.04	5.99	0.00	8.98	8.66	10.50	3.98	14.19	12.38	13.53
% false negative	1.12	1.19	0.00	1.23	1.77	1.95	0.93	0.75	0.69	0.19
% false positive	7.66	8.43	14.35	5.60	4.55	4.02	10.45	0.50	1.88	1.11
III. Germany (# manually labeled observations: 22,232)										
Parameter 1	0.74	−0.32	1.26	0.75	0.01	0.00	18.50	–	–	–
Parameter 2	–	–	–	15	–	–	–	–	–	–
Accuracy rank	6	7	10	5	8	9	4	3	2	1
% correct (median)	68.93	63.98	63.44	73.34	63.60	63.50	81.76	85.34	87.41	90.37
(Standard deviations)	(0.61)	(0.66)	(0.65)	(0.56)	(0.69)	(0.66)	(0.55)	(0.58)	(1.52)	(0.53)
of which cycling	60.30	62.10	63.44	56.42	63.60	63.50	58.00	58.09	58.42	59.08
of which not	8.63	1.88	0.00	16.92	0.00	0.00	23.76	27.25	28.99	31.28
% false negative	3.01	1.48	0.00	6.99	0.00	0.00	6.17	5.39	4.88	3.71
% false positive	28.06	34.54	36.56	19.66	36.40	36.50	12.07	9.27	7.71	5.93

Note: These alternative results are based only on "cleaner" subsamples that are either unanimously labeled as "cycling" by all RAs or not labeled as "cycling" by any RA. Other details follow Table 2.

Figure 4: Gains from Additional Data in "Cleaner" Datasets



Note: These alternative results are based only on "cleaner" (i.e., less ambiguous) subsamples that are either unanimously labeled as "cycling" by all RAs or not labeled as "cycling" by any RA.

Appendix D Materials for "Discussions" in Section 5

D.1. Why Existing Methods Work in Australia But Fail in Germany Figure 5 plots histograms of the median daily change of prices/margins, which underlies the simplest of the asymmetry-based methods (Method 3). Most of the manually identified cycles in Australia exhibit asymmetry (Panels I–II), whereas German cycles are not necessarily asymmetric (Panel III).

D.2. Why Margins Correlate Positively with Cycles in Australia But Negatively in Germany The scatter plots of Figure 6 show the means and the standard deviations of margins are positively correlated in all datasets. As the histograms of Figure 7 show, however, their volatility and (manually identified) cyclicality are correlated positively in Australia but negatively in Germany. Thus, margins and their cyclicality are correlated positively in Australia but negatively in Germany.

D.3. Why Do Existing Methods Find "Positive Correlations"? Figure 8 shows histograms of standard deviations by "cyclicality" based on Methods 3 and 4. These pictures suggest the threshold rules underlying the asymmetry-based methods tend to flag high-volatility cases as "cycles," because only sufficiently large movements can satisfy these conditions. As Figure 6 shows, however, high volatility is a poor predictor of true cyclicality (based on manual classification) in the German data.

D.4. Manual Benchmark Focuses on "Cyclicality" But Not "Asymmetry." Maybe "Asymmetric Cycles" Do Feature Higher Markups? Table 8 compares the mean differences of margins based on our manual benchmark (copied from column 0 of Table 3) and its refined version in which we further require "asymmetry" based on the negative median change, $median_{d\in t}\left(\Delta p_{i,d}\right) < 0$, as an additional criterion for (Edgeworth) cycles. Nevertheless, the results are similar both qualitatively and quantitatively.

Figure 5: Histograms of Median Daily Change

I. Western Australia **II. New South Wales** **III. Germany**

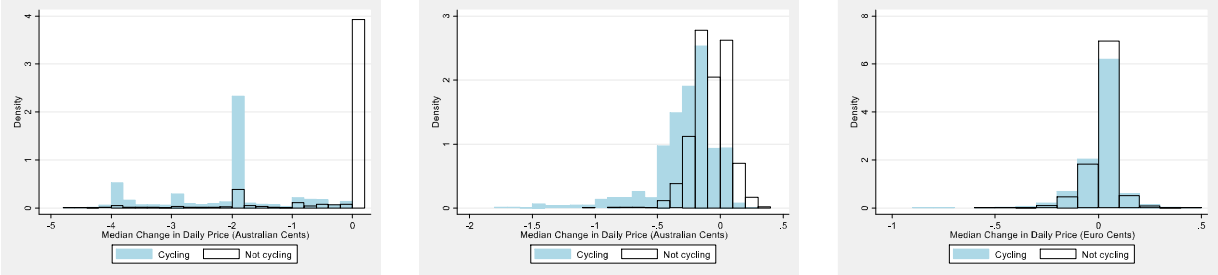


Figure 6: Scatter Plots of Mean and Standard Deviation

I. Western Australia **II. New South Wales** **III. Germany**

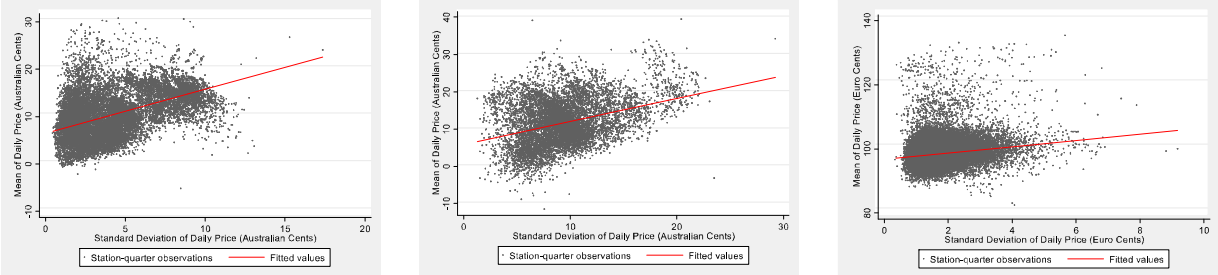


Figure 7: Histograms of Standard Deviation

I. Western Australia **II. New South Wales** **III. Germany**

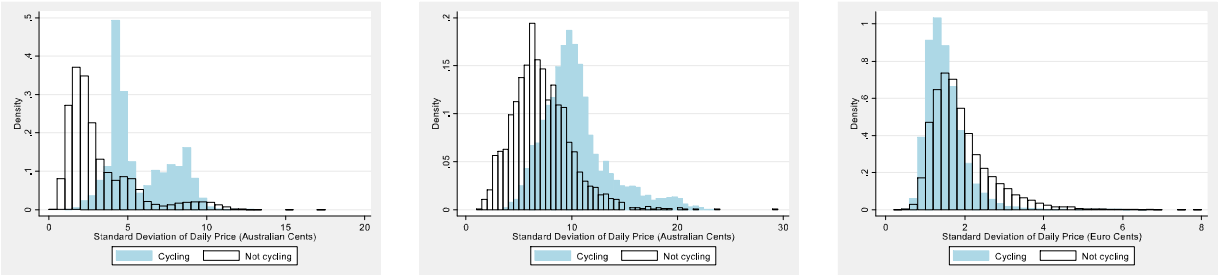


Figure 8: Asymmetry-based Definitions Pick Up Volatility

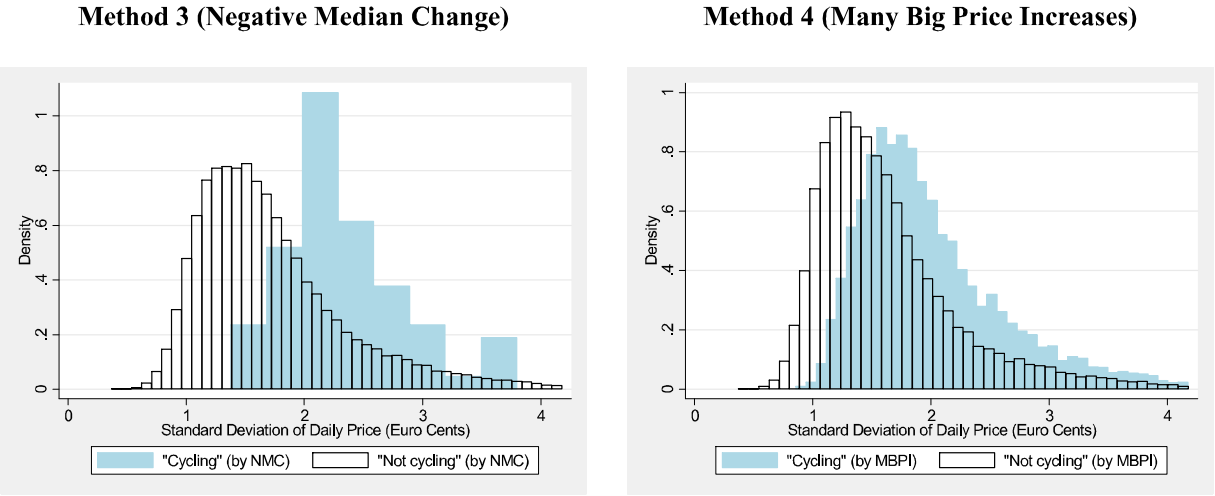


Table 8: Profit Margins by Cycle Status and Asymmetry

	(0)	(00)
Method	Manual	Manual + asymmetry
	III. Germany	
Cycling		
# obs.	14,116	4,265
Mean	98.18	98.01
Std. dev.	3.57	3.39
Not cycling		
# obs.	21,569	31,420
Mean	98.65	98.53
Std. dev.	4.37	4.16
Difference		
Mean diff.	−0.47	−0.52
Welch's t	−11.11	−9.05
D. F.	33,984	6,153
p value	$< .001$	$< .001$

Note: Column (0) is the same as in Table 3, which Column (00) refines by asymmetry based on negative median change. The unit of measurement (of means and standard deviations) is the euro cent. The p value indicates the probability that the difference in means is zero based on Welch's t statistic and the approximate degrees of freedom.

Appendix E Practitioner's Guide

We suggest the following steps for automating the detection of Edgeworth cycles:

1. Eyeball and manually categorize a random sample of 100 station-quarter observations in terms of cyclicality (but not necessarily asymmetry).

2. If both cyclical and non-cyclical cases exist, calibrate the two-parameter model of Method 4 (MBPI) as a first attempt to distinguish them algorithmically.

3. For more formal, mathematical definitions of cyclicality, use Methods 5 (FT) and 6 (LS), both of which are readily implementable in many programming languages for scientific computing.

4. If the accuracy of MBPI, FT, and LS is unsatisfactory, try Methods 7 (CS), 9 (E-RF, which nests Methods 1–7), 8 (LSTM), and 10 (E-LSTM, which nests Methods 1–8), in the increasing order of complexity and expected accuracy.

5. Once the detection of cyclicality (as recognized by humans) is successfully automated, refine the classification of "cycling" observations in terms of asymmetry. The median-price-change statistic from Method 3 (NMC) offers a simple way to capture asymmetry.[49] For example, one can distinguish between the Edgeworth-type asymmetry (i.e., the median change is negative), the inverse-Edgeworth asymmetry (i.e., the median change is positive), and symmetry (i.e., the median change is approximately zero).

6. If desired, this asymmetry-based classification can be automated by using some clustering algorithm on the distribution (e.g., a histogram) of the median price change across station-quarter observations.

7. Once the classification based on both cyclicality and asymmetry is complete, compute the mean margin and other statistics for each type of observations (e.g., Table 3). Welch's t statistic and the associated degrees of freedom can be used for testing the null hypothesis that the means of the two subsamples (of potentially different sizes) are equal.

8. The previous step assumes the dataset contains only prices and margins. If additional data are available on the characteristics of gas stations and their locations (as well as

[49] Methods 1 and 2 are slightly more complex but similar.

other demand- and supply-side factors such as competition), control for these additional covariates in a suitable regression model.

Thus, even though Method 10 (E-LSTM) is our top pick in terms of cycle-detection accuracy, other methods (including the existing ones) have important roles to play, both as a tool for initial inspection and as a summary statistic to characterize asymmetry.

References

[1] Assad, Stephanie, Robert Clark, Daniel Ershov, and Lei Xu. 2021. "Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market," working paper.

[2] Breiman, Leo. 2001. "Random forests," *Machine Learning*, 45: 5–32.

[3] Byrne, David P.. 2019. "Gasoline Pricing in the Country and the City," *Review of Industrial Organization*, 55: 209–235.

[4] Byrne, David P., and Nicolas de Roos. 2019. "Learning to Coordinate: A Study in Retail Gasoline," *American Economic Review*, 109 (2): 591–619.

[5] Byrne, David P., Jia Sheen Nah, and Peng Xue. 2018. "Australia Has the World's Best Petrol Price Data: FuelWatch and FuelCheck," *Australian Economic Review*, 51 (4): 564–577.

[6] Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. "An Empirical Comparison of Supervised Learning Algorithms," *Proceedings of the 23rd International Conference on Machine Learning*, 161–168.

[7] Castanias, Rick, and Herb Johnson. 1993. "Retail Gasoline Price Fluctuations," *Review of Economics and Statistics*, 75 (1): 171–174.

[8] Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models," in James Heckman and Edward Leamer, eds., *Handbook of Econometrics, Volume 6B*. Amsterdam, Netherlands: North Holland (Elsevier).

[9] Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv preprint*:1406.1078.

[10] Clark, Robert, and Jean-François Houde. 2014. "The Effect of Explicit Communication on pricing: Evidence from the Collapse of a Gasoline Cartel," *Journal of Industrial Economics*, 62 (2): 191–228.

[11] Deltas, George. 2008. "Retail Gasoline Price Dynamics and Local Market Power," *Journal of Industrial Economics*, 56 (3): 613–628.

[12] Doyle, Josephy, Erich Muehlegger, and Krislert Samphantharak. 2010. "Edgeworth cycles revisited," *Energy Economics*, 32 (3): 651–660.

[13] Eckert, Andrew. 2002. "Retail price cycles and response asymmetry," *Canadian Journal of Economics*, 35 (1): 52–77.

[14] Eckert, Andrew. 2003. "Retail price cycles and the presence of small firms," *International Journal of Industrial Organization*, 21: 151–170.

[15] Eckert, Andrew, and Heather Eckert. 2013. "Regional Patterns in Gasoline Station Rationalization in Canada," *Journal of Industry, Competition and Trade*, 14: 99–122.

[16] Edgeworth, Francis Ysidro. 1925. "The Pure Theory of Monopoly," in *Papers Relating to Political Economy, Vol. 1.* (London: MacMillan), pp. 111–142.

[17] Foros, Øystein, and Frode Steen. 2013. "Vertical Control and Price Cycles in Gasoline Retailing," *Scandinavian Journal of Economics*, 115 (3): 640–661.

[18] Greff, Klaus, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber. 2017. "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, 28 (10): 2222–2232.

[19] Hansen, Bruce E.. 2020. *Econometrics*, University of Wisconsin, manuscript.

[20] Haucap, Justus, Ulrich Heimeshoff, and Manuel Siekmann. 2015. "Fuel Prices and Station Heterogeneity on Retail Gasoline Markets," working paper.

[21] Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long short-term memory," *Neural Computation*, 9 (8): 1735–1780.

[22] Klein, Timo. 2021. "Autonomous algorithmic collusion: Q-learning under sequential pricing," *RAND Journal of Economics*, 52 (3): 538–599.

[23] Lewis, Matthew S.. 2009. "Temporary Wholesale Gasoline Price Spikes Have Long-Lasting Retail Effects: The Aftermath of Hurricane Rita," *Journal of Law and Economics*, 52: 581–605.

[24] Lewis, Matthew S.. 2012. "Price leadership and coordination in retail gasoline markets with price cycles," *International Journal of Industrial Organization*, 30: 342–351.

[25] Lewis, Matthew, and Michael Noel. 2011. "The Speed of Gasoline Price Response in Markets with and without Edgeworth Cycles," *Review of Economics and Statistics*, 93 (2): 672–682.

[26] Lomb, N. R.. 1976. "Least-squares frequency analysis of unequally spaced data," *Astrophysics and Space Science*, 39: 447–462.

[27] Martin, Simon. 2018. "Market Transparency and Consumer Search: Evidence from the German Retail Gasoline Market," working paper.

[28] Maskin, Eric, and Jean Tirole. 1988. "A Theory of Dynamic Oligopoly, II: Price Competition, Kinked Demand Curves, and Edgeworth Cycles," *Econometrica*, 56 (3): 571–599.

[29] Murphy, Kevin P.. 2012. *Machine Learning: A Probabilistic Perspective*, Cambridge, MA: The MIT Press.

[30] Musolff, Leon. 2021. "Algorithmic Pricing Facilitates Tacit Collusion: Evidence from E-Commerce," working paper, Princeton University.

[31] Noel, Michael D.. 2007. "Edgeworth Price Cycles: Evidence from the Toronto Retail Gasoline Market," *Journal of Industrial Economics*, 55 (1): 69–92.

[32] Noel, Michael D.. 2015. "Do Edgeworth price cycles lead to higher or lower prices?" *International Journal of Industrial Organization*, 42: 81–93.

[33] Noel, Michael D.. 2018. "Calendar synchronization of gasoline price increases," *Journal of Economics and Management Strategy*, 28: 355–370.

[34] Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C*, Second Edition. Cambridge, MA: Cambridge University Press.

[35] Scargle, Jeffrey D.. 1982. "Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data," *Astrophysical Journal*, 263: 835–853.

[36] VanderPlas, Jacob T.. 2018. "Understanding the Lomb–Scargle Periodogram," *Astrophysical Journal Supplement Series*, 236 (16): 1–28.

[37] Wang, Zhongmin. 2008. "Collusive Communication and Pricing Coordination in a Retail Gasoline Market," *Review of Industrial Organization*, 32: 35–52.

[38] Wills-Johnson, Nick, and Harry Bloch. 2010. "The Shape and Frequency of Edgeworth Price Cycles in an Australian Retail Gasoline Market." Unpublished working paper, Curtin University of Technology.

[39] Zimmerman, Paul R., John M. Yun, and Christopher T. Taylor. 2013. "Edgeworth Price Cycles in Gasoline: Evidence from the United States," *Review of Industrial Organization*, 42: 297–320.