Detecting Edgeworth Cycles*

Timothy Holt† Mitsuru Igami‡ Simon Scheidegger§

November 14, 2023

Abstract

We develop and test algorithms to detect "Edgeworth cycles," which are 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 the 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. We conclude with a set of practical recommendations.

Keywords: Deep neural networks, Edgeworth cycles, Fuel prices, Machine learning, Markups, Nonparametric methods, Spectral analysis

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, Cuicui Chen for detailed discussions, Robert Clark, Daniel Ershov, Simon Martin, and Felix Montag 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/conference participants at the Düsseldorf Institute for Competition Economics (DICE), Brown University, Cornell University, Yale University, IIOC 2022, and ES-NASM 2022 for comments. This work was supported by the Swiss National Science Foundation (SNF), under project ID "New methods for asset pricing with frictions." The replication package is available at https://dx.doi.org/10.5281/zenodo.10126406.

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

‡Department of Economics, Yale University. 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 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).[1]

These asymmetric movements are called Edgeworth cycles and have been studied extensively.[2] 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 that 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. Because theory provides only a loose characterization of Edgeworth cycles, empirical researchers have to rely on visual inspections and summary statistics based on a single quantifiable characteristic: asymmetry. Meanwhile, the phenomena's most basic property, cyclicality, is almost completely absent from the existing operational definitions. Even though asymmetry may be the most salient feature of—and hence a necessary condition for—Edgeworth cycles, it is not a sufficient condition. Empirical findings are only as good as the measures they employ; the incompleteness of detection methods could affect the reliability of "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 scalable detection methods represents an important practical

[1] 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).

[2] 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). We explain its theoretical background in section 2.

[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. The Bundeskartellamt does the

challenge for economists and policymakers.[4]

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.[5]

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 gasoline 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.[6] Given the lack of a clear theoretical definition, we construct a benchmark "ground truth" based on human recognition of price cycles as follows. We reorganize the raw data as panel data of the daily margins (= retail minus wholesale prices) of gasoline stations 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.[7] We employ eight 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.[8] Note that we look only for cyclicality and do not impose asymmetry or other criteria in the manual-classification stage. The reason is that asymmetry is—unlike cyclicality—amenable to clear

same in Germany.

[4]Systematic methods to detect price cycles are useful for researchers who do *not* want to study cycles as well. Chandra and Tappata (2011) examine the role of consumer search in generating temporal dispersion in the US retail gasoline prices. However, they could not completely reject Edgeworth cycles as an alternative explanation (see their page 697 and footnote 46) because they did not have a scalable method to prove the absence of cycles in their large dataset of more than 25,000 stations. Our procedure would have allowed them to provide more concrete evidence.

[5]Section 4 formally introduces all methods.

[6]See Byrne, Nah, and Xue (2018) for a guide to the Australian data. Haucap, Heimeshoff, and Siekmann (2017), Martin (2018), and Assad, Clark, Ershov, and Xu (2021), among others, study the German data.

[7]We explain our data, the choice of sampling frequency, and the manual classification procedures in section 3.

[8]Appendix sections B.4–B.6 show results under alternative criteria.

mathematical definitions and can easily be checked at a later stage. Hence, we prioritize the detection of cyclicality, thereby alleviating the cognitive burden on RAs.

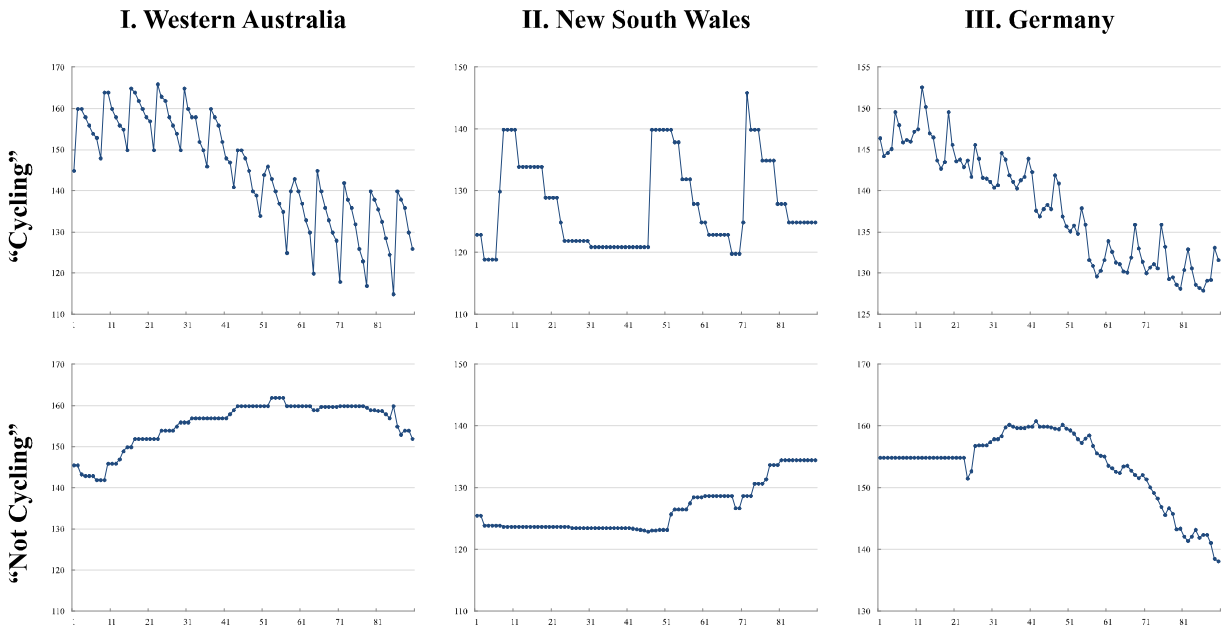
At this point, one might wonder whether human recognition of cycles is an appropriate benchmark. We regard it as the best *feasible* option (the "second best") given the lack of clear theoretical definitions (the "first best"). Manual classification by a team of RAs represents a best-effort practice in the literature and provides a relevant—if not perfect—benchmark in the following sense. First, most existing studies employ some rule-of-thumb definitions with calibrated thresholds, which are ultimately based on the researchers' eyeballing and judgment, the details of which are rarely documented. We make such procedures more explicit, systematic, and transparent, so that the overall scheme becomes more reproducible. Second, human recognition is central to the prominence of Edgeworth cycles as an antitrust topic. Despite the lack of universal definitions, the phenomena have become a perennial policy issue in many countries precisely because consumers and politicians can easily recognize cyclical patterns when they see them. In this regard, human recognition *is* the "ground truth" that eventually determines the phenomena's relevance to public policy. We interpret our RAs' responses as a proxy for the general public's responses to various patterns in gasoline prices.

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, are 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 because our interview with a German industry expert suggests Edgeworth cycles are known to exist. They are (in fact) called the "price parachute" (or *Preis Fallschirm*) phenomena, and are considered to be part of common pricing strategies among practitioners. The Bundeskartellamt (2011) also confirms the existence of both weekly and daily cycles.[9] Methods 7–10 attain 71%–80% accuracy even in this challenging environment.

Second, we assess the cost effectiveness of each method by using only 0.1%, 1%, 5%, 10%, $\cdots$, 80% 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 performance, but their data requirement is

[9] See sections 3.3 and 7.2 for further details on the German data.

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



Note: The top panels and the bottom panels show examples of daily retail-price series in "cycling" and "non-cycling" station-quarter observations, respectively, for illustration purposes. The vertical axes measure retail gasoline prices in the Australian cent (left, middle) and the Euro cent (right), respectively. The horizontal axes represent calendar days. Note our main analysis uses retail margins (= retail minus wholesale prices) instead, thereby controlling for costs.

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 gasoline 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). This finding emerges under both "cyclicality only" and "cyclicality with asymmetry" definitions of Edgeworth cycles. 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.

The rest of the paper is organized as follows. Sections 2–4 explain the theoretical background, data, and methods, respectively. Sections 5–7 report our main findings and discuss their economic/policy implications. Section 8 summarizes our practical recommendations for cycle detection. Section 9 concludes.

Related Literature, Contributions, and Replication Package. This work is so closely connected to the Edgeworth-cycle literature and cites so many related works throughout the paper that a separate review section would be redundant. Specifically, the first five paragraphs of this introductory section provide the overall literature context; section 2 covers the theoretical background; section 3 cites data sources as well as several papers that use the German data; section 4.1 acknowledges the proponents of each of the existing methods; section 4.2 suggests helpful readings for the new methods that we propose.

Besides the contributions specific to the phenomena, our broader contribution is threefold: (i) introducing certain "heavy-duty" machine-learning models and methods (a class of deep-neural-network architectures) to the empirical economics literature, (ii) precisely explaining the mechanisms inside these "black boxes," and (iii) demonstrating their usefulness with a concrete, public-policy-relevant example.

For the purpose of lowering the "entry barriers" for those empirical economists who are considering the use of advanced machine-learning tools, we have made the computer code (in Python), the dataset, and detailed documentations (including the read-me file and the Online Appendix) publicly available as a replication package (Holt, Igami, and Scheidegger 2023) at https://dx.doi.org/10.5281/zenodo.10126406.

2 Theoretical Background

Even though the primary goal of this article is empirical, some conceptual anchoring clarifies the target of measurement.

2.1 What Are Edgeworth Cycles?

Maskin and Tirole (1988) offer the following verbal description: "In the Edgeworth cycle story, firms undercut each other successively to increase their market share (price war phase) until the war becomes too costly, at which point some firm increases its price. The other firms then follow suit (relenting phase), after which price cutting begins again. The market price thus evolves in cycles" (pages 571–572). This description and its micro foundation—as a class of Markov perfect equilibria (MPE) in an alternating-move dynamic duopoly game—suggest four important characteristics: cyclicality, asymmetry, stochasticity, and strategicness.

Property 1: Cyclicality. The price should exhibit cyclicality, as the terminology suggests. However, this property is not so obvious in Edgeworth's (1925) original conjecture. His writing focuses on the indeterminacy of static equilibrium in a price-setting game between capacity-constrained duopolists. Even though he mentions a price path that resembles Maskin and Tirole's description as an example, he uses the word "cycle" only once. More generally, he conjectures that "there will be an indeterminate tract through which the index of value will oscillate, or rather will vibrate irregularly for an indefinite length of time" (page 118). Thus, Edgeworth's original theory features not so much cyclicality as "perpetual motion" (page 121).

Nevertheless, we have chosen to focus on cyclicality in this paper. Theoretically, Maskin and Tirole's equilibrium strategies (their equation 23) explicitly feature price cycles. Empirically, it is this repetitive pattern that draws consumers' and politicians' attention; "perpetual motion" alone would not raise antitrust concerns.

Property 2: Asymmetry. The second characteristic is the asymmetry between relatively few large price increases and many small price decreases. Edgeworth (1925) does not emphasize this property either, but it plays an important role in the Maskin-Tirole formalization and the subsequent empirical literature (see Methods 1–4 in section 4.1).

Property 3: Stochasticity. In Maskin and Tirole's Edgeworth-cycle MPE, big price increases are supposed to happen stochastically, not deterministically. The reason is that if one firm always "relents" whenever the low price is reached, the other firm will always wait and free-ride, which in turn would make the first firm more cautious about the timing of price increases. Thus, the frequency of cycles must be stochastic—with varying lengths of

time spent at the low price—in equilibrium.[10] We do not impose stochastic frequencies as a necessary condition in our empirical procedures, but some of our methods are designed to accommodate cycles with varying frequencies (Methods 7 and 8 in section 4.2).

Property 4: Strategicness. The cyclical patterns are supposed to emerge from dynamic strategic interactions between oligopolistic firms. If similar patterns are observed under monopoly, their underlying mechanism must be different from that of Edgeworth cycles.[11] Thus, whether market structure is monopolistic or oligopolistic is a theoretically important distinction. Empirically, however, market definition is rarely clear-cut in practice. Even when a gasoline station is located in a geographically isolated place, pricing decisions at large chains tend to be centralized at the city, region, or country level. Market structure at these aggregate levels is oligopolistic in all of our datasets. Consequently, we do not impose any geographical boundaries a priori. We simply analyze data at the individual station level.[12] Our idea is that once the station-level characterization is successfully completed, one can always compare cyclicality across stations in the same market (defined geographically or otherwise) and look for synchronicity—whenever such analysis becomes necessary.

2.2 Are Edgeworth Cycles Competitive or Collusive?

Whether Edgeworth cycles represent collusion is a subtle issue on which we do not take a stand. Several reasons contribute to its subtlety and our cautious attitude.

First, the theoretical literature seems agnostic about the distinction between competitive and collusive behaviors in the current context. On the one hand, Edgeworth's (1925) narrative lacks any hint of cooperative actions or intentions. On the other hand, Maskin and Tirole (1988) seem open to collusive interpretations: "Several of the results of this paper underscore the relatively high profits that firms can earn when the discount factor is near 1. Thus our model can be viewed as a theory of tacit collusion" (page 592). In the more recent literature, however, the term "tacit collusion" is usually associated with collusive equilibria

[10] This theoretical property seems largely overlooked in the empirical literature, presumably because the first two properties make the phenomena sufficiently interesting and policy-relevant.

[11] Alternative explanations include consumers with heterogeneous search costs, intertemporal price discrimination, and "dynamic pricing" algorithms (broadly defined as any pricing strategy and its implementation(s) that tries to exploit consumer heterogeneity and time-varying price-elasticity of demand).

[12] This operational decision is not without its own risks. For example, if the grid of relevant prices were very coarse and two firms take turns to change prices, we might not be able to observe clear cycles at any specific station's time-series data even if such cycles exist at the aggregate level. Fortunately, gasoline prices reside on a relatively fine grid with the minimum interval of the Australian or Euro cent. Moreover, Maskin and Tirole's Edgeworth-cycle MPE requires a fine grid with sufficiently small intervals (denoted by k in their model). Therefore, we believe the risk of missing aggregate cycles is low.

in repeated-games models.[13] The latter rely on the concepts of monitoring, punishment, and history-dependent strategies as their underlying mechanism, none of which are prominently featured in Edgeworth cycles. Thus, even though Maskin and Tirole's own remarks suggest the possibility of collusive interpretations, we feel inclined to regard their Edgeworth-cycle MPE as a reflection of competitive interaction between forward-looking oligopolists.

Second, in terms of antitrust law, explicit communications of a cooperative nature are the single most important act that constitutes criminal price-fixing. That is, tacit collusion is not illegal as long as it truly lacks explicit communication. Notwithstanding this legal distinction, most of the theoretical literature does not discriminate between tacit and explicit collusion because the process through which firms reach collusive agreements is usually not modeled. Hence, an important gap lies between economic theory and legal enforcement, which complicates the interpretation of Edgeworth cycles in empirical research.

Third, partly reflecting this unresolved theory-enforcement divide, the empirical literature has documented many different instances of asymmetric price cycles, both *with* and *without* legally established evidence of criminal price-fixing. Accordingly, interpretations of observed cycles vary across papers on a case-by-case basis. The only common thread that unites the large empirical literature is the data patterns with clear cyclicality and asymmetry.

For these reasons, we do not (necessarily) interpret Edgeworth cycles as evidence of collusion. Consequently, we do not aim or claim to detect "collusion." Reliable methods to detect price cycles would nevertheless be useful for detecting cycle-based collusion.

3 Data and Manual Classification

Retail-price data are publicly available for the universe of individual gasoline stations in WA, NSW, and Germany. 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 the retail price,

$$p_{i,d} \equiv p^R_{i,d} - p^W_{i,d}, \tag{1}$$

[13] Tirole and his coauthors exclusively focus on the repeated-games theory when they summarize the "economics of tacit collusion" for the European competition authority. See Ivaldi, Jullien, Rey, Seabright, and Tirole (2003).

where $p^R_{i,d}$ and $p^W_{i,d}$ are retail and wholesale prices at station i on day d, and simply refer to this markup measure $p_{i,d}$ 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.

3.1 Data Sources and Preparation

Retail Prices. We use three datasets on retail gasoline prices that are publicly available and of high quality. *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.[14] The Market Transparency Unit for Fuels of the *Bundeskartellamt* publishes similar data for every German gas station in minute intervals.[15]

Sampling Frequencies. The raw data from WA contain daily retail 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 (see section 3.3 for further details on Germany).

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.[16]

3.2 Manual-Classification Procedures

Whereas most existing studies treat the manual-verification process as an informal preparatory step (to be embodied by the analyst's eventual choice of methods and calibration of threshold parameters), we make it as systematic as possible. Our goal is to develop and

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

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

[16] Regional wholesale prices are the most detailed publicly available information on the operating costs of retail gasoline stations (to our knowledge). We do not observe station-specific costs.

compare the performance of multiple methods, and such "horse racing" requires a common benchmark.

To establish a "ground truth" based on human recognition of cycles, we employed a team of eight RAs to manually classify station-quarter observations.[17] 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). The manual labeling of the datasets proceeded in three stages.

WA. 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 to maintain consistency. As a result, each station-quarter (i, t) in WA has one label based on the consensus of the two RAs.

NSW. 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 agreed with the coauthor's at least 80% of the time. Subsequently, each of the RAs independently labeled the entire dataset in February–April 2021. Thus, each (i, t) in NSW carries three labels.

Germany. 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.

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

Risk of "Collusion" Is Low. In the computerized procedures for NSW and Germany, each RA is given *one randomly selected observation for labeling* at a time. We believe the risk of "collusion" among RAs is low because copying each other's answers would require (i) keeping records of random sequences of thousands of observations with their station-quarter identifiers, (ii) exchanging these long records, and (iii) matching each other's answers across different random sequences. Such a conspiracy is conceivable in principle but prohibitively time-consuming in practice. Honestly labeling all observations just once would be much easier.

Summary Statistics. 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 decisions), and 0 otherwise.[18] Thus, we prepare the target for automatic detection in a relatively conservative manner.

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 main text for details.

[18] We assess the sensitivity of our results under alternative criteria in Appendix sections B.4–B.6.

3.3 Rationale for Daily Frequency and Quarterly Window

Several considerations led us to use the daily sampling frequency and the quarterly time window.

First, we prioritize setting a common time frame for all three datasets. Our goal is to compare the performance of various methods in multiple different datasets under the same protocol; a detailed case study of any single region/country is not our main objective. The daily frequency is the finest granularity that can be commonly used across all datasets because retail prices in WA are fixed for 24 hours due to regulation (see section 3.1). It is also the finest granularity used in most other studies (however, see below for our discussion of the German data).

Second, cyclicality implies repetition, the identification of which requires a sufficiently long time window. The existing studies on WA and NSW report cycles with frequencies of one to several weeks, whereas those on Germany report both weekly and intra-day cycles. The 12–13 weeks of a calendar quarter permit repeated observations of relatively long (e.g., monthly) cycles.

Third, shorter-than-daily (e.g., hourly) frequencies would be too "costly" for our research design, as systematic manual verification is its essential component. Eyeballing and labeling a 10% subsample of the entire German dataset at the hourly (instead of daily) frequency would require 24 times more labor: 480 hours $\times$ 24 = 11,520 hours. At the hourly wage of \$13.50, the total cost would be \$155,520.

Fourth, we avoid longer-than-quarterly time windows for two reasons. One is that macroeconomic factors (such as business cycles, financial crises, and geopolitical upheavals in the world crude oil market) tend to feature prominently in a time horizon longer than 90 days, which increases noise. Another reason is that longer windows tend to complicate classification, as cycles might appear in only one part of the graph but not others.

For these reasons, the daily frequency and the quarterly horizon are suitable for our purposes. Note that our choice is driven by the comparative research design, practical considerations, and budget constraints, not conceptual limitations. All of the methods can be applied to time-series data of any frequency and length in principle.

On Intra-Day Cycles in the German Data. We are aware of multiple studies that document intra-day price cycles in Germany. The first investigation into the German retail fuel markets by Bundeskartellamt (2011) studies data from four major cities (Hamburg, Leipzig, Cologne, and Munich) in January 2007–June 2010 and highlights three patterns.

First, weekly cycles exist in both diesel and gasoline prices, with the highest prices on Fridays and the lowest prices on Sundays and Mondays. Second, intra-day cycles exist as well, with many small price reductions during the day and fewer, larger increases in the evening. Third, stations operated by Aral (BP) and Shell typically lead those price increases, in which one follows the other within three hours in 90% of the cases, followed by three other major chains.

Given the well-documented presence of intra-day cycles, one might wonder whether our focus on the daily data and multi-day cycles leads to an important omission. Our answer is "yes," but this issue is orthogonal to the main purpose of this research.

By aggregating the underlying minute-by-minute data to 24-hour averages, we lose these interesting short-run movements. Our choice of the daily frequency is driven by the comparative design of our research, which prioritizes the systematic comparisons across the three datasets and (costly) manual verification. Thus, researchers who wish to conduct an in-depth case study of the German fuel markets might want to analyze intra-day patterns as well.

Nevertheless, the presence of shorter cycles does not preclude that of longer cycles; Bundeskartellamt (2011) confirms the existence of both (see above). One should also note that the intra-day cycles seem to follow a specific time schedule in which prices (i) rapidly increase at night between 20:00 and 24:00 hours and (ii) gradually decrease from around 6:00 in the following morning (Siekmann 2017). As Linder (2018) correctly points out, such a deterministic pattern is more consistent with intertemporal price discrimination than Maskin and Tirole's Edgeworth cycles (recall Property 3—stochasticity—in section 2.1). Hence, while interesting, the intra-day cycles in Germany are outside the scope of this paper.

4 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.

4.1 Existing Methods Mostly Focus on Asymmetry

The existing methods in the literature almost exclusively focus on asymmetry. We formalize four of them as simple parametric models.

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

$$mean\left(len\left(run^{+}\right)\right) < mean\left(len\left(run^{-}\right)\right) + \theta^{PRNR}, \tag{2}$$

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.[19]

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

$$\left|mean_{d\in t}\left(\Delta p_{i,d}^{+}\right)\right| > \left|mean_{d\in t}\left(\Delta p_{i,d}^{-}\right)\right| + \theta^{MIMD}, \tag{3}$$

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.[20]

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

$$median_{d\in t}\left(\Delta p_{i,d}\right) < \theta^{NMC}, \tag{4}$$

where $\Delta p_{i,d}$ denotes a 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.[21]

[19] 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.

[20] 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.

[21] Many subsequent studies use this method, 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.

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

$$\sum_{d \in t} \mathbb{I}\left\{\Delta p_{i,d} > \theta_1^{MBPI}\right\} \geq \theta_2^{MBPI}, \tag{5}$$

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.[22] Thus, many instances of big price increases are taken as evidence of price cycles.

Other Existing Methods. These methods are among the most cited in the literature, but our listing 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.[23] 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. However, they are not specifically designed for defining or detecting cycles.

4.2 Our Proposals to Capture Cyclicality

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.

[22] Lewis (2009) also uses a similar method, with $\theta_1^{MBPI} = 4$ (US cents/gallon) in a single day or two consecutive days.

[23] 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 probability 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.

This subsection is rather technical because we are introducing data-analysis techniques from outside the usual toolbox of empirical economists. If the reader is not interested in methodological details, a quick look at the first and the last few sentences of each method would be sufficient for an overview. If, instead, the reader wants to exactly follow our procedures, Appendix A.1 (and the replication package) provides additional 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

$$P(f) \equiv \frac{1}{N} \left| \sum_{n=1}^{N} g_n e^{-2\pi i f x_n} \right|^2, \tag{6}$$

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 $(g_n)_{n=1}^{N}$.[24]

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

$$\max_f P_{i,t}(f) > \theta_{\max}^{FT}, \tag{7}$$

where $P_{i,t}(f)$ is the periodogram (6) of station-quarter (i,t), and $\theta_{\max}^{FT} > 0$ 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.[25] 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

[24] Appendix A.1 (Method 5) introduces FT to readers who are not familiar with Fourier analysis.

[25] 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 A.1 (Method 6). Another is practical: its off-the-shelf computational implementation can offer more granular periodograms.

classical periodogram (6):[26]

$$P^{LS}(f) = \frac{1}{2}\left\{ \frac{\left(\sum_n g_n \cos\left(2\pi f\left[x_n - \tau\right]\right)\right)^2}{\sum_n \cos^2\left(2\pi f\left[x_n - \tau\right]\right)} + \frac{\left(\sum_n g_n \sin\left(2\pi f\left[x_n - \tau\right]\right)\right)^2}{\sum_n \sin^2\left(2\pi f\left[x_n - \tau\right]\right)} \right\}, \tag{8}$$

where τ is specified for each frequency f as

$$\tau = \frac{1}{4\pi f}\tan^{-1}\left(\frac{\sum_n \sin\left(4\pi f x_n\right)}{\sum_n \cos\left(4\pi f x_n\right)}\right). \tag{9}$$

We propose the following threshold condition to detect cycles:

$$\max_f P_{i,t}^{LS}(f) > \theta_{\max}^{LS}, \tag{10}$$

where $\theta_{\max}^{LS} > 0$ 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). That is, we smooth the discrete (daily) time series by interpolating it with a commonly used continuous function.[27] For each (i,t), we fit CS to its demeaned price series, $\overline{p}_{i,d} \equiv p_{i,d} - mean_{d\in t}\left(p_{i,d}\right)$, and count the number of times the fitted function $\overline{CS}_{i,t}(d)$ crosses the d-axis (i.e., equals 0). Operationally, we count the number of real roots and detect cycles with the condition,

$$\#roots\left(\overline{CS}_{i,t}(d)\right) > \theta_{root}^{CS}, \tag{11}$$

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 network (ANN) 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.

[26] Appendix A.1 (Method 6) explains how this expression relates to FT.

[27] We use a cubic Hermite interpolater, which is a spline where each piece is a third-degree polynomial of Hermite form. Appendix A.1 (Method 7) explains the details of this functional form.

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 use a multi-layer architecture (a.k.a. "deep" neural networks) to enhance the model's flexibility.[28] 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:

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

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

where $d = 1, 2, \cdots, D$ is our index of days, $\Delta p_d \equiv p_d - p_{d-1}$ (we set $\Delta p_1 = 0$), $\tanh(x) \equiv \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is the hyperbolic tangent function, $\circ$ denotes the Hadamard (element-wise) product, and $\Lambda(x) \equiv \frac{e^x}{1+e^x}$ is the cumulative distribution function (CDF) of the logistic distribution.[29] 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

$$s^*\left(\mathbf{p}_{i,t}; \boldsymbol{\theta}^{LSTM}\right) \equiv \omega_{10} + \boldsymbol{\omega}_{11}' \mathbf{s}_D^L > 0, \tag{14}$$

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

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

[29] See Appendix A.1 (Method 8) for further details on this specification and computational implementation.

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 the value of $\boldsymbol{\omega}$ that approximately maximizes the accuracy of prediction (to be explained in section 4.3 and Appendix A.2).

In summary, LSTM sequentially processes the daily price data in a flexible Markov model with many latent states, 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

$$g_{i,t}^m \equiv LHS_{i,t}^m - RHS_{i,t}^m \tag{15}$$

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 (3) defines Method 2. Hence, $g_{i,t}^2 = \left| mean_{d \in t} \left(\Delta p_{i,d}^+ \right) \right| - \left| mean_{d \in t} \left(\Delta p_{i,d}^- \right) \right|$.[30] Let

$$\mathbf{g}_{i,t} \equiv \left(g_{i,t}^m \right)_{m=1}^M \tag{16}$$

denote their vector, where $M = 7$.[31] We construct a decision-tree classification algorithm that takes $\mathbf{g}_{i,t}$ as inputs and predicts $cycle_{i,t} = 1$ if and only if

$$h\left(\mathbf{g}_{i,t}; \boldsymbol{\omega}^{RF}, \boldsymbol{\kappa}^{RF} \right) \equiv \sum_{k=1}^K \omega_k^{RF} \mathbb{I}\left\{ \mathbf{g}_{i,t} \in R_k \right\} \equiv \sum_{k=1}^K \omega_k^{RF} \phi\left(\mathbf{g}_{i,t}; \boldsymbol{\kappa}_k^{RF} \right) > 0, \tag{17}$$

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.[32] Because finding the truly optimal partitioning is a computationally difficult (combinatorial) problem, we use an RF algorithm to stochastically approximate it.[33] Thus, this method aggregates and generalizes Methods 1–7 in a flexible manner that permits (i) mul-

[30] 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}.

[31] Our computational implementation also incorporates two additional variants of each of Methods 5–7, which we explain in Appendix A.1 (Method 9). Hence, the eventual value of M is $7 + (2 \times 3) = 13$.

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

[33] See Appendix A.1 (Method 9) for further details.

tiple 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 (16) as additional variables in the laws of motion:

$$\begin{aligned} \mathbf{s}_d^l &= \tanh\left(\mathbf{c}_d^l\right) \circ \Lambda\left(\boldsymbol{\omega}_1^l + \boldsymbol{\omega}_2^l \Delta p_d + \boldsymbol{\omega}_3^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{12}^l \mathbf{g}\right), \text{ and} \quad (18) \\ \mathbf{c}_d^l &= \tanh\left(\boldsymbol{\omega}_4^l + \boldsymbol{\omega}_5^l \Delta p_d + \boldsymbol{\omega}_6^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{13}^l \mathbf{g}\right) \circ \Lambda\left(\boldsymbol{\omega}_7^l + \boldsymbol{\omega}_8^l \Delta p_d + \boldsymbol{\omega}_9^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{14}^l \mathbf{g}\right) \\ &\quad + \mathbf{c}_d^{l-1} \circ \left[1 - \Lambda\left(\boldsymbol{\omega}_7^l + \boldsymbol{\omega}_8^l \Delta p_d + \boldsymbol{\omega}_9^l \mathbf{s}_d^{l-1} + \boldsymbol{\omega}_{14}^l \mathbf{g}\right)\right], \quad (19) \end{aligned}$$

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.

4.3 Optimization of Parameter Values ("Training")

Accuracy Maximization. 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,

$$\%\text{ correct}\left(\boldsymbol{\theta}\right) \equiv \frac{\sum_{(i,t)} \mathbb{I}\left\{\widehat{cycle}_{i,t}\left(\boldsymbol{\theta}\right) = cycle_{i,t}\right\}}{\#\ all\ predictions} \times 100, \quad (20)$$

where $\widehat{cycle}_{i,t}\left(\boldsymbol{\theta}\right) \in \{0,1\}$ 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 A.2. Thus,

$$\boldsymbol{\theta}^* \equiv \arg\max_{\boldsymbol{\theta}} \quad \%\text{ correct}\left(\boldsymbol{\theta}\right) \quad (21)$$

characterizes the optimized (or "trained") model for each method.[34]

Splitting Data into Training and Testing Subsamples. We optimize and evaluate each method as follows, separately for each of the three datasets (WA, NSW, and Germany):

1. Randomly split each labeled dataset into an 80% "training" subsample and a 20% "testing" subsample.

[34] See Appendix A.2 for further details.

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.[35]

4. Repeat these three steps 101 times.[36]

5. Report the medians of the optimized parameter values, as well as the medians and standard deviations of the prediction-accuracy results.

5 Results

Table 2 summarizes the performance of all methods for each dataset. We report the median accuracy, the composition of correct and incorrect predictions, and the associated parameter value(s), $\boldsymbol{\theta}^*$, for each method.

WA. Panel I shows the results in WA, where clear-cut cycles of deterministic frequencies are known to exist. Almost all methods achieve high accuracy near or above 90%. The flexible, nonparametric models of Methods 8–10 do particularly well with above 99% accuracy.

Some of the parameter values are informative about the underlying data patterns. For example, CS lags behind all other methods with (a still respectable) 85% accuracy. Its parameter value, $\theta^{CS}_{roots} = 22.5$, suggests the model is trained to focus on shorter cycles with wavelengths 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, the inferior performance of CS stems from missing the latter, longer cycles.

Another interesting result concerns MBPI, which achieves 90% accuracy. Byrne and de Roos (2019) set $\theta^{MBPI}_1 = 6$ and $\theta^{MBPI}_2 = 3.75$ in their original study of WA. Our accuracy-maximizing values (5.05 and 5, respectively) turn out to be reasonably close to their calibrated values. This comparison illustrates how experienced researchers' parameter tuning could approximate the results of systematic numerical optimization. One can also interpret this finding as an external validation of our manual classification. Given the similar parameter values and the high accuracy, it follows that our manual classification must be broadly consistent with Byrne and de Roos's eyeballing results.

[35] 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.

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

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

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%). This finding alone does not necessarily invalidate the use of NMC in other datasets but cautions against overly relying on any single metric.

Germany. 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).[37] 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.

This profile of success and failure is intriguing. The methods that exclusively focus on asymmetry (Methods 1–3) and deterministic cycles (Methods 5–6) fail, whereas those that capture cyclicality in "fuzzier" manners (Methods 4 and 7) manage to make at least some correct (non-degenerate) predictions. These results suggest that not all of the German cycles conform to the idealized patterns of asymmetry or cyclicality and that less rigid classification rules could be relatively more robust to irregular patterns and noise.

The parameter values of CS ($\theta^{CS}_{roots} = 24.50$) and MBPI ($\theta^{MBPI}_{2} = 14$) suggest that the German cycles are approximately weekly. That is, $\theta^{CS}_{roots} = 24.50$ means at least as many ups and downs are often recorded in "cycling" observations, which translate into the wavelength of $90 \div \frac{24.5}{2} = 7.3$ days or shorter. Likewise, $\theta^{MBPI}_{2} = 14$ requires at least as many "big jumps" within a calendar quarter and hence implies the wavelength of $90 \div 14 = 6.4$ days or shorter. These numbers provide another opportunity for external validation: the detailed case study by Bundeskartellamt (2011) confirms the presence of weekly cycles (see section

[37] As 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 B.4 reports results based on "cleaner" subsamples that eliminate such observations with disagreements. By contrast, Appendix B.5 investigates how the algorithms classify ambiguous observations (i.e., station-quarters on which RAs disagree and/or choose "maybe cycling"). Appendix B.6 examines such labeler heterogeneity in detail.

3.3).

Summary. 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 performance. 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).

Performance on Simulated Cycles. In Appendix A.3, we examine the 10 methods' performances on simulated data with four types of artificial patterns: white noise, theoretical Edgeworth cycles, "reverse Edgeworth" cycles, and sine waves of various lengths. We simulate 10,000 quarters of data based on each DGP and deploy the three pre-trained versions (WA, NSW, and Germany) of each method. Four findings emerge. First, Methods 1–4 and 7 either fail to detect most of these cycles or incorrectly classify white noise as cycles. Second, Methods 5–6 are the best performers in such a controlled environment. Third, the performances of Methods 8–10 are somewhere between these two groups of methods. Fourth, a little bit of additional noise could either help or hinder the performance of these 10 methods. These results suggest the real-world data are qualitatively different from simulated data with artificial cycles.

6 How Much Data Do We Need?

The accuracy "horse racing" in the previous section shows that more flexible methods tend to outperform simple parametric ones, which is not surprising. The real question is the cost of "training" complicated machine-learning algorithms, which are known to require a lot of data. This section investigates the cost-accuracy trade-offs of the 10 methods.

The accuracy of cycle detection naturally improves with the size of the training dataset. The rate of improvement is different across methods, however. Figure 2 shows performance

when we restrict the training dataset to only 0.1%, 1%, 5%, 10%, $\cdots$, 80% 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 cost of human-generated labels is negligible for such small samples (US$3.51, US$2.84, and US$6.48, respectively, based on the hourly wage of US$13.50 for undergraduate RA work at Yale University as of 2021). These methods are extremely cost effective.

The fact that simple models with one or two parameters (Methods 1–7) require only a few dozen observations is not surprising. All we have to do is to adjust one or two numerical thresholds to distinguish cycles from non-cycles. However, the finding that E-RF (Method 9) is equally cheap *is* surprising. It is a highly nonlinear machine-learning model with potentially many thresholds and interactions. This result suggests that the building blocks of E-RF—the summary statistics derived from Methods 1–7—contain genuinely useful information that those stand-alone methods under-utilize.

Methods 8 and 10 contain a few thousand parameters and obviously need more data. For instance, E-LSTM's accuracy in NSW is below 50% when it uses only 10 observations (0.1% subsamples). Fortunately, their performance dramatically improves with a mere 1% subsample, and they start outperforming all other methods when 5% subsamples are used.[38] 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.[39] Thus, even though LSTM and E-LSTM require more data for a given accuracy level, their total cost is surprisingly low, making them the highest-accuracy methods within a limited amount of resources.

This finding is unexpected, but is definitely good news: heavy-duty machine-learning algorithms turn out to be not only useful, but also affordable in the context of detecting Edgeworth cycles. Our conjecture is that the cyclical patterns that humans recognize are relatively simple after all, even though explicitly articulating them might be difficult.

[38] 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 11 and 12 in Appendix B.3).

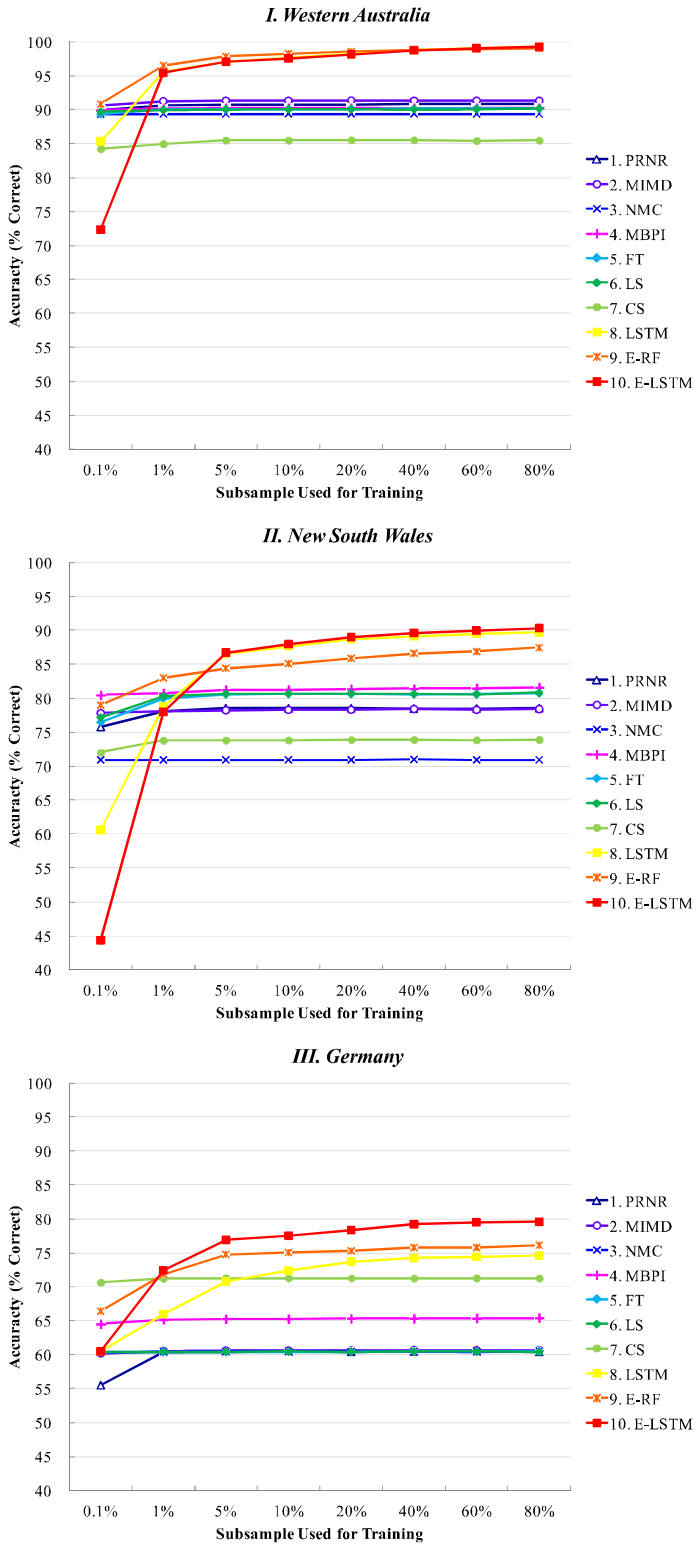
[39] Panel (B) of Table 11 in Appendix B.3 reports the total cost of manual labeling for each dateset. The reason only "several hundred observations" are sufficient to approximately optimize "a few thousand parameters" is because various forms of regularization restrict the effective parameter space.

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 4 for the definition of each method. Appendix B.1 investigates whether combining some or all of Methods 1–4 leads to better performances. Appendix B.2 reports additional results for the variants of Methods 5–7. 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. The replication package (Holt, Igami, and Scheidegger 2023) implements the training and testing of each method only once, but adding a loop in the computer code (for 101 repetitions) should be straightforward.

Figure 2: Gains from Additional Data



Note: The exact numbers underlying these plots are reported in Panel (A) of Table 11 in Appendix B.3.

7 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 amounts of data on fuel markets. Some papers find that the presence of cycles is positively correlated with retail prices and markups, whereas others find the opposite relationships.[40] Section 7.1 investigates how such findings depend on the definition of cycles. Sections 7.2 and 7.3 report additional findings.

7.1 Cycles and Margins

Human-Recognized Cyclicality and Margins. Table 3 compares the retail-wholesale margins between "cycling" and "non-cycling" observations.[41] 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.[42]

Algorithmic Cycle Detection and Margins. 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 also emerge in NSW, even though one method fails (Method 3) and one reaches the opposite conclusion (Method 7). These discrepancies suggest that researchers find a positive or negative cycle-margin relationship depending 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

[40] The former includes Deltas (2008), Clark and Houde (2014), and Byrne (2019); the latter includes Lewis (2009), Zimmerman et al. (2013), and Noel (2015).

[41] Our measure of profit margin is the difference between the retail price and the wholesale price before tax, as defined in equation (1) in section 3, in the Australian cent in WA and NSW and the euro cent in Germany, respectively. Note the lack of volume data—a main limitation in this area of research—means that we cannot check the extent to which consumers buy at the bottom of price cycles.

[42] Determining the exact source of heterogeneity is beyond the scope of this paper. There can be many reasons and Edgeworth cycles are only one of the possible mechanisms. Our purpose is to illustrate with concrete examples how different methods could lead to different findings and policy implications.

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 5), but Method 4 features reasonable parameter values and achieves at least 65% accuracy. Hence, we cannot dismiss these discrepancies as purely random anomalies.

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.

Margins and *Asymmetric* Cycles. Note that our classification so far has focused on cyclicality but not asymmetry. One might wonder whether our findings could change if we study *asymmetric* cycles specifically. The answer is "no." The results are virtually the same when we focus on asymmetric cycles.

Table 4: 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.

Table 4 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. The results are similar both qualitatively and quantitatively.

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

7.2 Additional Findings

The results in sections 5, 6, and 7.1 constitute our main findings, but the curious patterns in section 7.1 present additional puzzles. We address them in the following and report supporting evidence in Appendix C.

1. Why Existing Methods Work in Australia But Fail in Germany. Most of the cycles in Australia follow specific (almost deterministic) frequencies and exhibit strong asym-

metry, whereas German cycles are noisier and not always asymmetric (see supplementary plots in Appendix C.1). The existence of asymmetric *non*-cycles in Germany further complicates the issue. Hence, asymmetry-based methods correctly identify cycles in Australia but not in Germany.

2. 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. That is, higher markups tend to accompany higher volatility. The reason is that retail and wholesale prices are relatively close so that the only direction in which margins can move significantly 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 C.2 for supplementary plots). Therefore, the average level and cyclicality of margins are correlated positively in Australia but negatively in Germany.

3. 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, which explains the existence of "volatile non-cycles" in the data.

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

5. Could Intra-Day Cycles Be the Source of Curious Patterns in Germany? The answer is "yes" and "no." In general, our daily sampling frequency and 90-day window are suitable for identifying cycles with the frequencies of several days to a month or so. Shorter frequencies may not be well represented.

Nevertheless, if the "intra-day" cycles follow the frequency of exactly 24 hours (or any hours that can divide 24 evenly), they would be "averaged out" in the process of computing daily prices and would not affect our observations. The existing studies suggest that they do follow exactly 24-hour cycles (see section 3.3). Hence, how intra-day cycles affect the multi-daily volatility in our data is not obvious.[43]

[43] One possibility is the existence of "medium frequency" cycles that are longer than 24 hours, but shorter than 3–4 days. However, we are not aware of any studies that document such cycles. In short, the coexistence of daily, weekly, and other cycles and their interactions constitute an open-ended question for further research.

6. Why Manual Classification Provides a Relevant Benchmark. At this point, one might question (again) the relevance of human recognition as a benchmark. Our answer is still the same as in section 1 (paragraph 6): It is the "second best" option. If we had a perfect mathematical definition, no detection problem would arise in the first place. In the absence of such a formula, the existing research relied on rules of thumb that were ultimately validated by selective eyeballing by the authors. We made this process more systematic and transparent.

7.3 Exploratory Data Analysis

As a further demonstration of the use of automatic cycle detection, this section investigates the distribution of price cycles across time and space. Obviously, such an exploratory data analysis becomes possible only after a scalable method to detect cycles is used on the entire dataset. We first describe time-series patterns and then explore cross-sectional correlations.

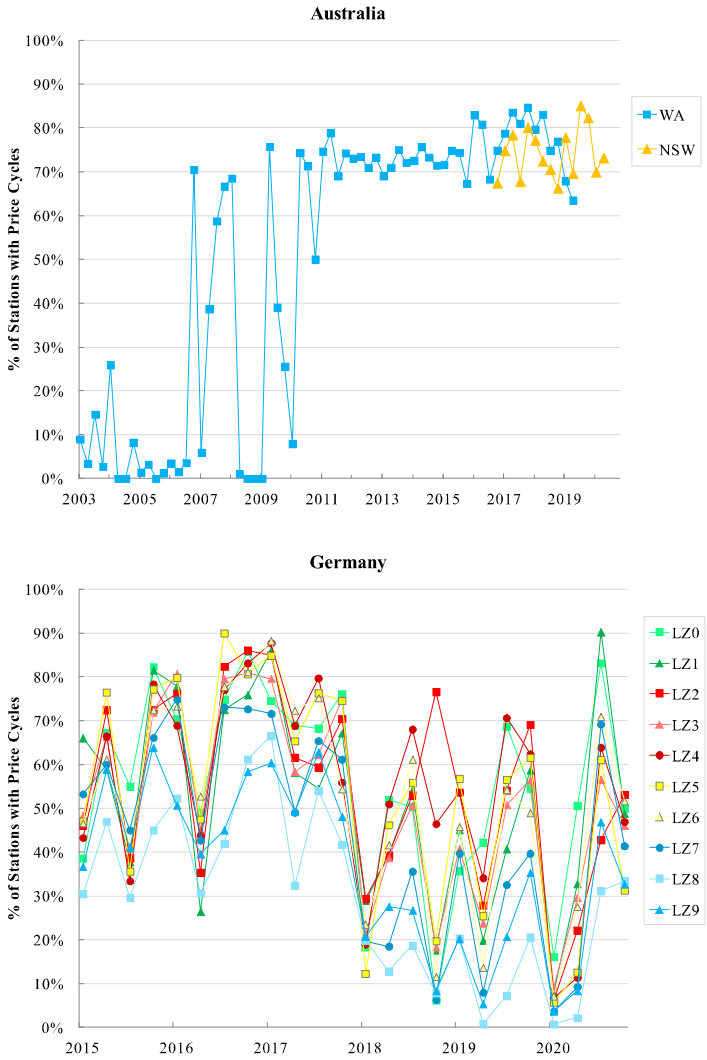
Time Series Patterns. How many stations exhibit price cycles at each point in time? The two panels of Figure 3 plot the fractions of stations that exhibit price cycles in Australia and Germany, respectively. Throughout this section, the recognition of cycles is based on the median-performance version (parameter values) of the most accurate algorithm (Method 10), which we apply to the entire dataset—both labeled and unlabeled—in each region/country.

The two regions of Australia, WA and NSW, show mostly high percentages of cycling stations. WA offers the longest data period. Byrne and de Roos (2019) documented clear price cycles in two subperiods (2007–2008 and 2010–2015), both of which correspond to the periods in which cycles are prevalent according to our method.[44] Thus, the results of our method confirm Byrne and de Roos's description of the WA data in terms of time series. The NSW dataset starts relatively recently in 2016:Q4. Its range of approximately 70%–90% is comparable to WA.

The German picture is more "colorful," with greater heterogeneity across regions. We show the fraction of cycling stations in each of the 10 geographic zones (*Postleitzonen*, henceforth LZs). LZ0 and LZ1 (in green) correspond to former East Germany; LZ2–LZ6 (in red and yellow) are northwestern regions; LZ7–LZ9 (in blue) roughly correspond to the southern

[44] Readers might wonder what causes sudden increases and decreases in WA in the 2000s. Some of them reflect genuine changes in the number of cycling stations; others could be due to noise in the original data because the WA database lacks a consistent station identifier. Even though we tried to reconstruct as "balanced" panel data as possible (based on street addresses and other observable characteristics), the recorded number of stations varies across time, sometimes quite dramatically.

Figure 3: How Many Stations Exhibit Price Cycles?



Note: LZ0–LZ9 are Germany's 10 geographic zones (*Postleitzonen*). See main text for details.

states of Baden-Württemberg and Bavaria.[45] Three patterns emerge. First, whereas LZ0–LZ6 tend to move together in relatively high ranges, LZ7–LZ9 exhibit consistently lower percentages. Second, despite these differences in levels, all regions display similar fluctuations most of the time, and such fluctuations could be large. Third, as a general trend, the overall range shifted downward from 30%–90% in 2015–2017 to 0%–70% in 2018–2019. The timing of this change would seem to roughly coincide with the introduction and dissemina-

[45]For maps and further details, see Wikipedia page on "Postal codes in Germany" at https://en.wikipedia.org/wiki/Postal_codes_in_Germany (accessed on January 10, 2023).

tion of automatic pricing algorithms (see Assad et al. (2021)), but the clarification of their causal relationship would require further research.

In Appendix D.1, we also investigate the relationship between macroeconomic shocks and price cycles, which seems complex.

Spatial Patterns. We now explore spatial patterns within each region. The geographical scope of price cycles (e.g., local, city-wide, or regional) and their synchronization patterns might shed light on their mechanism and potentially inform the definition of relevant markets for antitrust purposes.

Specifically, we investigate whether multiple gasoline stations tend to exhibit price cycles at the same time, and if so, how such tendencies change with the distance between them.

We construct our measure of "correlation" between stations as follows. First, within each region, we list all possible pairs of stations and split them into seven distance bins (less than 1km, 1–5km, ..., 50–100km, and above 100km) based on their Euclidean distances.[46] Second, for each pair, we calculate the *percentage of quarters in which their cycle statuses match* (i.e., either both stations exhibit cycles or neither of them does). Third, for each distance bin in each region, we take the average of these percentages, either across all pairs or across pairs of same-brand stations. This procedure creates a summary statistic of how well the presence or absence of cycles is synchronized across multiple stations in each region—and how their "correlation" varies with distance.[47]

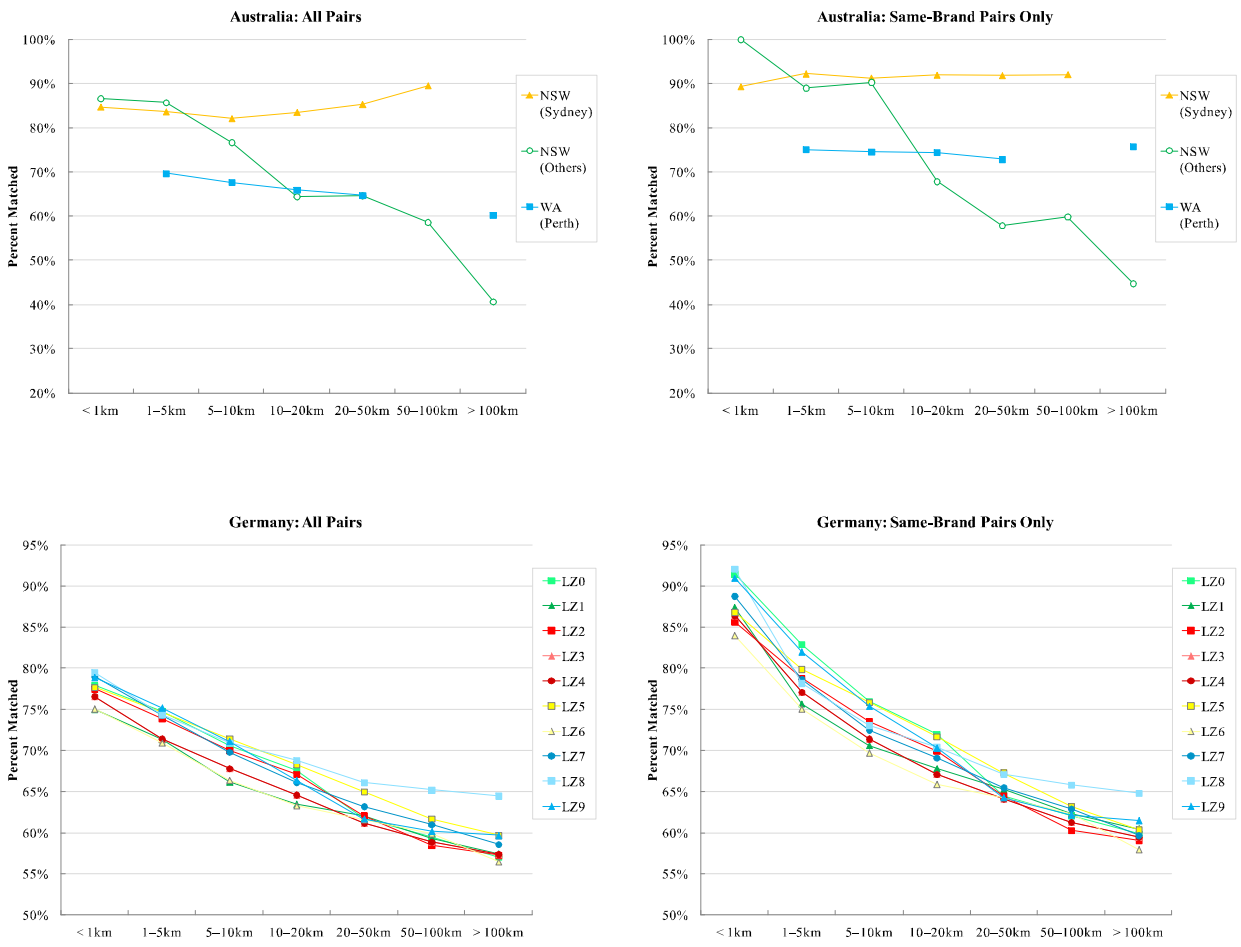
Figure 4 reports the spatial patterns of "correlation" in four graphs: (i) all pairs in WA and NSW, (ii) same-brand pairs in WA and NSW, (iii) all pairs in Germany, and (iv) same-brand pairs in Germany. Four patterns emerge. First, the majority of the station-pair-quarter observations shares cycle status, with the exception of the most distant (>100 km) bin in rural NSW. Second, the cities and the rural areas of Australia exhibit qualitatively different patterns. The station pairs within Perth and Sydney (the capital cities of WA and NSW, respectively) tend to show high correlations with limited variability across distance bins, whereas the rest of NSW features "correlations" that decrease with distance.[48] Third, all 10 LZs of Germany show similar patterns in which "correlations" steadily decrease with distance. Fourth, pairs of same-brand stations tend to be more correlated than all/any pairs in both Australia and Germany, especially in the 0km–10km bins in Germany.

[46] Note we consider only pairs that share at least 12 calendar quarters of valid data in common.

[47] We say "correlation" in quotes because we use the "percentage of quarters with matched cycle statuses" instead of correlation coefficient, which is undefined when a station always (or never) shows cycles.

[48] All of the WA stations (with sufficient observations for these plots) are in Perth, which is why we do not split WA into urban and rural areas as in NSW.

Figure 4: How Presence of Cycles Correlates between Stations



Note: See main text for the definition of "percent matched" statistics. The WA graphs do not show markers or lines in two distance bins (less than 1km and 50–100km) because the WA data have relatively few pairs in these bins, which we grouped in the adjacent bins for the purpose of visualization.

In Appendix D.2, we further investigate how the relationship between cycle correlation and distance varies by time period in WA.

8 Practical Recommendations

Based on our findings in sections 5–7, we suggest the following steps as a practical (but not necessarily the most rigorous) guide for automating the detection of Edgeworth cycles:

1. Choose the data frequency and time window that would permit the identification of hypothesized cycles. That is, the sampling frequency must be shorter than that of

suspected cycles, and the time horizon should accommodate at least a few repetitions. (For the sake of simple exposition, our explanation in the following keeps assuming the daily frequency and the quarterly window.)

2. Eyeball and manually categorize a random sample of 100 station-quarter observations in terms of cyclicality (but not necessarily asymmetry).[49] If sufficient numbers of both cyclical and non-cyclical cases are found, proceed to the next step. If not, increase the sample size.

3. As a first attempt to algorithmically distinguish cycles from non-cycles, calibrate one of the simpler methods. We recommend the two-parameter model of Method 4 (MBPI) because it is the only one (among Methods 1–4) that captures the notion of cyclicality.

4. For more formal, mathematical definitions of cyclicality, use Methods 5 (FT) or 6 (LS), both of which are readily implementable in many programming languages for scientific computing. Method 7 (CS) is another option with similarly off-the-shelf implementations.

5. If the performance of these methods is unsatisfactory, try Methods 9 (E-RF), 8 (LSTM), and 10 (E-LSTM), in increasing order of complexity and expected accuracy.

6. 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. 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). Methods 1 (PRNR) and 2 (MIMD) can be used for the same purpose.

7. 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. This process can be designed as either supervised or unsupervised machine-learning tasks.

[49] Adversarial circumstances, such as antitrust cases, could potentially introduce biases in the manual labeling of data. Hence, the selection and training of human labelers (in more formal contexts that the one assumed here) might have to be treated with the same care as in the selection and training of jury in trials. Appendix E discusses this issue in detail.

8. Once the classification based on both cyclicality and asymmetry is complete, compute the mean margin and other statistics for each type of observation (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.

9. The previous step assumes that the dataset contains only prices and margins. If additional data are available on the characteristics of gasoline stations and their locations (as well as other demand- and supply-side factors such as competition), control for these additional covariates in a suitable regression model.

10. At any point after step 4, one might also consider another refinement based on the frequency of cycles. Cycles of multiple lengths may coexist within a single dataset (see sections 3.3 and 7.2). Methods 5–7 would be useful for this purpose.

Thus, even though Method 10 (E-LSTM) is the top runner 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 for refinement.

9 Conclusion

We propose scalable methods to detect Edgeworth cycles so that the growing amount of "big data" on fuel prices can be scrutinized. The failure of the existing methods in noisy data suggests further investigation would benefit from distinguishing "cyclicality" from "asymmetry." Our nonparametric methods achieve the highest accuracy; such flexible models typically require large amounts of training data, but the requirement is minimal in this context. Whether researchers discover a positive or negative statistical relationship between markups and cycles depends on the choice of method. Because such "facts" are supposed to inform regulations and competition policy, these methodological considerations are directly policy relevant.

Data/Code Availability. The replication package (Holt, Igami, and Scheidegger 2023) and the Online Appendix are publicly available at https://dx.doi.org/10.5281/zenodo.10126406.

Appendix A Methodological Details and Simulations

A.1 Details of the New Methods

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

$$G(f) \equiv \int_{-\infty}^{\infty} g(x) e^{-2\pi i f x} dx. \tag{22}$$

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 consisting of a weighted sum of the Dirac delta functions at $\pm f_0$.[50] 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,

$$P_g \equiv |\mathcal{F}\{g\}|^2, \tag{23}$$

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 $g(x)$.

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 (22):

$$G_{obs}(f) = \sum_{n=-\infty}^{\infty} g(n\Delta x) e^{-2\pi i f n \Delta x}. \tag{24}$$

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 $g_n \equiv g(n\Delta x)$ and $G_k \equiv G_{obs}(k\Delta f)$. Then, the sample analog of (22) is

$$G_k = \sum_{n=0}^{N} g_n e^{-2\pi i k n/N}. \tag{25}$$

[50] The Dirac delta function is $\delta(f) \equiv \int_{-\infty}^{\infty} e^{-2\pi i f x} dx$, and hence, we can write $\mathcal{F}\left\{e^{2\pi f_0 x}\right\} = \delta(f - f_0)$. 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: $\mathcal{F}\{\cos(2\pi f_0 x)\} = \frac{1}{2}[\delta(f - f_0) + \delta(f + f_0)]$ and $\mathcal{F}\{\sin(2\pi f_0 x)\} = \frac{1}{2i}[\delta(f - f_0) + \delta(f + f_0)]$. See VanderPlas (2018) for further details.

One can construct the sample analog of the Fourier power spectrum (23) as (6) in the main text. This is the "classical" or "Schuster" periodogram.[51]

A potential drawback of the threshold rule in (7) 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 "significant" cycles:

$$HHI_{i,t} \equiv \sum_{f} \left(\frac{P_{i,t}(f)}{\sum_f P_{i,t}(f)} \right)^2 > \theta_{hhi}^{FT}, \tag{26}$$

where $\theta_{hhi}^{FT} \in (0, 1]$ is a scalar threshold parameter.[52] 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 (6) appears different from (8), (6) 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 (6) and (8) is the denominators in (8).

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:

$$\hat{g}(x; f) = A_f \sin\left(2\pi f \left(x - \phi_f\right)\right), \tag{27}$$

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

$$SSR^{LS}(f) \equiv \sum_n \left(g_n - \hat{g}(x_n; f)\right)^2. \tag{28}$$

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

[52] 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.

Scargle (1982) shows the following periodogram is identical to (8):

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

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 $\tilde{P}^{LS}(f)$.

The HHI variant of the LS method is

$$HHI_{i,t}^{LS} \equiv \sum_f \left(\frac{P_{i,t}^{LS}(f)}{\sum_f P_{i,t}^{LS}(f)}\right)^2 > \theta_{hhi}^{LS}. \tag{29}$$

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

$$S_K(x) = \sum_{j=0}^{P} \beta_j x^j + \sum_{k=1}^{N} \beta_{P+k}(x - \tau_k)^P \mathbb{I}\{x \geq \tau_k\}, \tag{30}$$

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 $(\tau_1 < \tau_2 < \cdots < \tau_N)$.[53] It is a special case of a sieve/series approximation that constitutes a class of nonparametric regression methods.[54] 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).[55]

In addition to the indicator of frequent oscillations in (11), 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 $d \in [1, 90]$.

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

[54] 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.

[55] 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 set $cycle_{i,t} = 1$ if and only if

$$\int_1^{90} \underline{CS}_{i,t}(d) > \theta_{int}^{CS}, \tag{31}$$

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.

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

$$\sum_{d=1}^{90} |\bar{p}_{i,d}| > \theta_{abs}^{CS}, \tag{32}$$

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 $\mathbf{c}_d^l$ (13) 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 (12) or inside tanh and Λ in (13). 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.

A.2 Parameter Optimization

We define two types of prediction errors as follows:

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

$$\%\text{ false positive}\,(\boldsymbol{\theta}) \equiv \frac{\sum_{(i,t)} \mathbb{I}\left\{\widehat{cycle}_{i,t}\,(\boldsymbol{\theta}) = 1, cycle_{i,t} = 0\right\}}{\#\ all\ predictions} \times 100. \tag{34}$$

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 $\widehat{cycle}_{i,t}\,(\boldsymbol{\theta}) = 1$ or $\widehat{cycle}_{i,t}\,(\boldsymbol{\theta}) = 0$ for all (i,t), and hence are mostly irrelevant for the purpose of finding well-performing $\boldsymbol{\theta}$s.

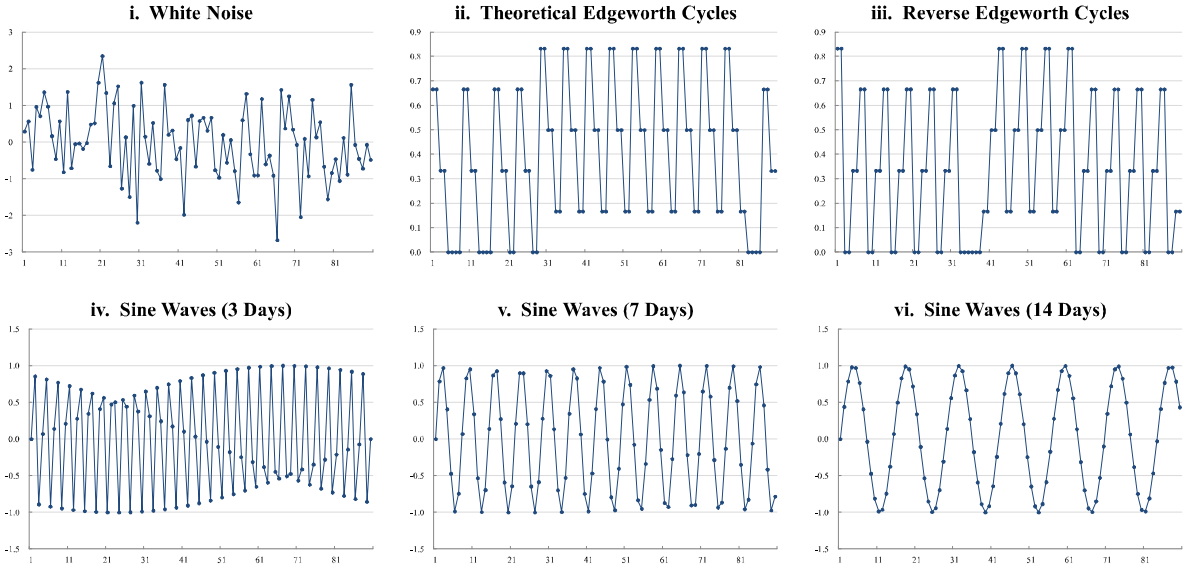
A.3 Performance on Simulated Cycles

This section studies the performance of each method on simulated data. Because artificial cycles and real-world cycles are qualitatively different, our purpose is not so much testing algorithms as understanding them better in a controlled environment.

Setup. We consider four kinds of DGPs: (i) white noise, (ii) Edgeworth cycles, (iii) "reverse Edgeworth" cycles, and (iv) sine waves of various lengths. First, white noise is simply

a 90-day sequence of i.i.d. random draws from the standard normal distribution. If a method "detects" cycles in white noise, we interpret it as a false positive. Second, we simulate theoretical Edgeworth cycles by using Maskin and Tirole's (1988) example.[56] Third, we simulate cycles with the opposite asymmetry (i.e., few big *decreases* and many small increases) by reversing the time stamps of simulated Edgeworth cycles in the above. Fourth, we generate sine waves (i.e., symmetric cycles) of five different wavelengths: 3, 7, 14, 21, and 28 days. We generate 10,000 quarters of simulated data based on each of these eight DGPs. Figure 5 shows examples of simulated cycles.

Figure 5: Examples of Simulated Cycles



Note: These pictures show examples of simulated price series before we transform them to match the mean and standard deviation of each dataset. The horizontal axes represent calendar days.

[56] We set the annual discount factor to 0.9, which translates into the daily discount factor of $\delta = 0.9997$ (in Maskin and Tirole's notation). The probability of a big price increase at the bottom of a price cycle (in their notation) is

$$\alpha(\delta) = \frac{\left(3\delta^2 - 1\right)\left(1 + \delta^2 + \delta^4\right)}{8 + 7\delta^2 + 2\delta^4 + 3\delta^6} \approx 0.2997.$$

With two firms taking turns to change prices in the grid of seven different price levels, $\left\{0, \frac{1}{6}, \frac{2}{6}, \ldots, 1\right\}$, the Edgeworth-cycle MPE entails asymmetric cycles of approximately weekly frequency. We simulate each 90-day sequence of data by randomly drawing one of the seven price levels as firm 2's initial price, to which firm 1 best-responds on day 1, to which firm 2 best-responds on day 2, and so on. These best responses are based on the equilibrium strategy profile of Maskin and Tirole's Table II. We use firm 1's prices as simulated data.

Each cycle-detection method comes in three versions as we optimize its parameters in three different datasets: WA, NSW, and Germany (Table 2). To match the mean and standard deviation of each (real) dataset, we rescale the simulated data through an affine transformation.

Results. Table 5 reports the percentages of simulated data (10,000 quarters each) that are classified as "cycling." Ideally, white noise should be classified as non-cycles and the rest as cycles, but that is not always the case. The top panel of Table 5 shows Method 4 (MBPI) and 7 (CS) mistakenly identify cycles in 100% of the white noise simulations. The reason is that they rely on counting the number of upward (and downward) price movements. White noise could trivially satisfy these criteria.

By contrast, Methods 1–3 are not fooled by white noise but fail to detect cycles in most other simulations, with the exception of Method 2 (MIMD) in theoretical Edgeworth cycles (53%). It is not surprising that these asymmetry-based methods fail to detect non-Edgeworth cycles; they are designed in such a way. Howerve, the result that Methods 1 (PRNR) and 3 (NMC) detect 0% of Edgeworth cycles *is* surprising. These false negatives are caused by the particular way in which Maskin and Tirole specify their model—each firm changes its price once every two days. There cannot be consecutive days of (strictly) positive or (strictly) negative "runs," which could fool Method 1. Likewise, Method 3 could be fooled by the 45 days of inaction in each 90-day sequence because the median price change is zero.

The "best" methods in the top panel are Methods 5 (FT) and 6 (LS) in the sense that they correctly reject most of the white noise as non-cycles and correctly detect most of the artificial cycles. In particular, their unique ability to detect cycles of any length is noteworthy. Even though they are trained in the WA data, which contain only weekly or two-week cycles, they correctly flag 100% of sine waves of both higher and lower frequencies.[57]

Finally, the performance of the nonparametric/machine-learning models of Methods 8 (LSTM), 9 (E-RF), and 10 (E-LSTM) are somewhere in the middle. On the one hand, they correctly reject 100% of the white noise as non-cycles and correctly detect cycles in some of the simulated cycles (theoretical Edgeworth cycles and weekly sine waves). On the other hand, they ignore most of the reverse Edgeworth cycles and the other sine waves. In other words, they faithfully detect patterns that they are trained to recognize (i.e., the cycles with Edgeworth-type asymmetry or approximately weekly frequency in the WA data) and reject others.

[57] We should note, of course, that these spectral methods are specifically designed for sine waves. Their real-world performance may not be as good, as we show in the main text.

Table 5: Performance on Simulated Data (% Classified as Cycling)

Method	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
I. All Models Trained with Labeled Data from Western Australia										
White noise	0	0	36	100	25	19	100	0	0	0
Edgeworth	0	53	0	100	91	92	100	28	90	40
Reverse Edgeworth	0	0	0	70	91	92	100	0	6	0
Sine wave: 3 days	0	0	0	100	100	100	100	0	0	0
Sine wave: 7 days	0	0	0	85	100	100	100	72	3	54
Sine wave: 14 days	0	0	0	12	100	100	0	23	0	26
Sine wave: 21 days	0	0	0	0	100	100	0	0	0	0
Sine wave: 28 days	100	0	0	0	100	100	0	0	0	0
II. All Models Trained with Labeled Data from New South Wales										
White noise	100	0	80	100	22	15	100	0	3	0
Edgeworth	100	100	100	100	98	99	100	0	29	0
Reverse Edgeworth	65	0	100	15	98	99	100	0	0	0
Sine wave: 3 days	100	0	100	100	100	100	100	0	0	0
Sine wave: 7 days	100	0	100	20	100	100	100	0	27	0
Sine wave: 14 days	100	0	100	0	100	100	100	0	100	0
Sine wave: 21 days	100	0	100	0	100	100	100	87	100	100
Sine wave: 28 days	100	0	100	0	100	100	100	100	100	47
III. All Models Trained with Labeled Data from Germany										
White noise	0	24	17	100	0	0	100	85	4	100
Edgeworth	0	100	0	64	0	0	100	86	49	100
Reverse Edgeworth	0	0	0	100	0	0	100	96	4	18
Sine wave: 3 days	0	0	0	100	0	0	100	56	8	100
Sine wave: 7 days	0	0	0	100	0	4	100	36	0	100
Sine wave: 14 days	0	0	0	100	0	4	0	0	0	4
Sine wave: 21 days	0	0	0	92	0	4	0	0	0	0
Sine wave: 28 days	0	0	0	44	0	4	0	0	0	0

Note: Each result (%) is based on 10,000 quarters of simulated data. See text for details.

The middle panel of Table 5 shows broadly similar results with the NSW-trained models. The original NSW data feature longer cycles of 2–4 week frequencies. Consequently, the nonparametric methods (8–10) respond only to the sine waves of relatively long wavelengths. The spectral methods (5 and 6) do well across all DGPs. Most of the existing methods (and Method 7) make degenerate predictions, but Method 2 happens to make perfectly correct predictions in the first two DGPs (white noise and Edgeworth cycles).

The bottom panel of Table 5 reports the performances of the models trained in the German data, in which cycles are noisy, nuanced, and generally difficult to detect. Almost all methods produce degenerate predictions on simulated data, with the exception of Method

2 (in the first two DGPs). Method 9 is another exception. Methods 4, 7, 8, and 10 make useless predictions by classifying most or all of white noise as cycles, but they are capable of distinguishing between shorter and longer cycles (i.e., they respond to shorter cycles but not longer ones). This distinction reflects the fact that typical cycles in Germany are weekly.

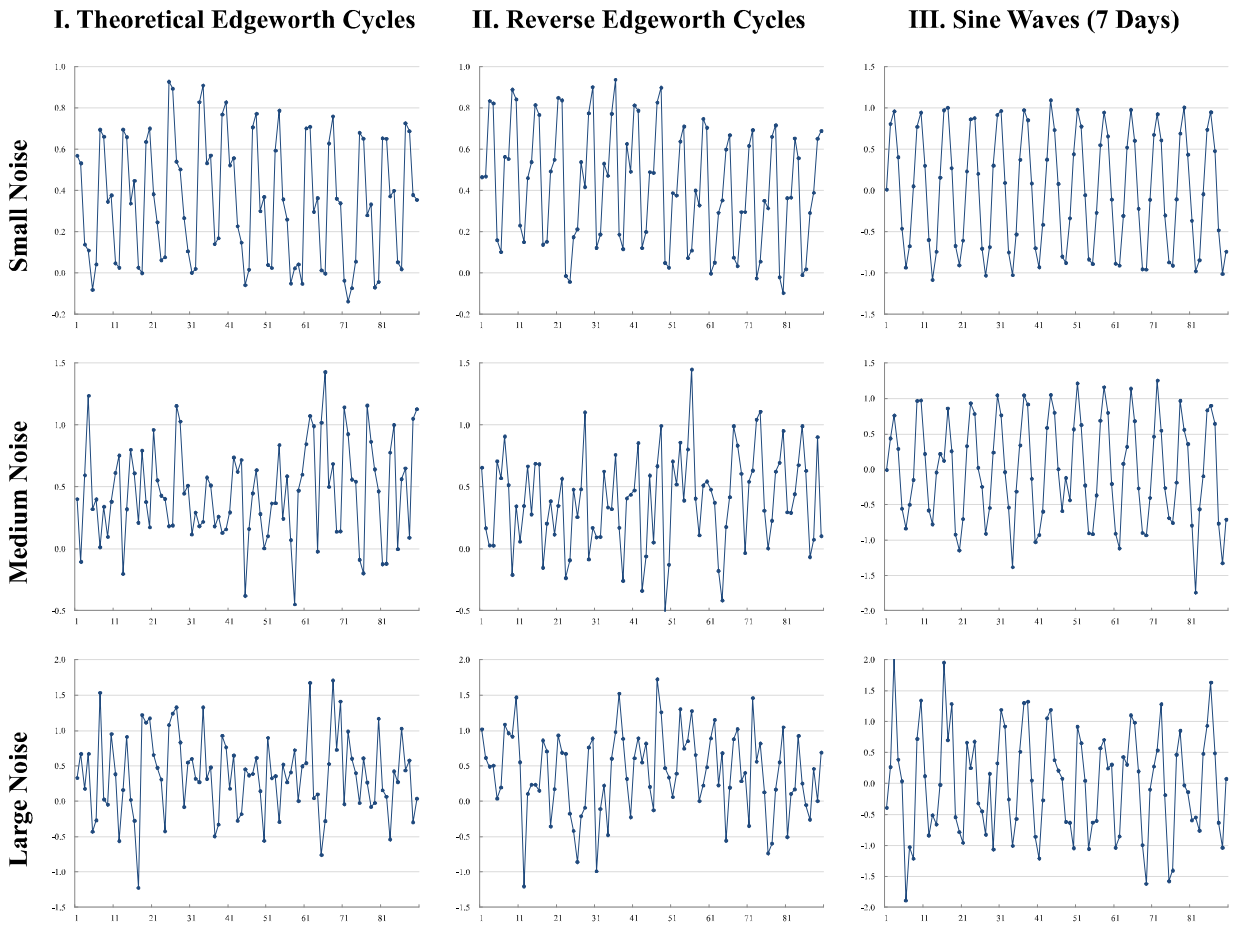
In conclusion, these simulations further clarify the performance characteristics of the algorithms. Certain methods (mostly Methods 1–4 and 7) struggle to reject white noise, whereas the spectral methods (Methods 5 and 6) perform well across the board—as long as they are trained on the data with clear cycles (WA and NSW). More flexible models (Methods 8–10) adapt to nuanced, specific data patterns. The 10 methods' relative performance rankings in the simulated data are often radically different from the ones in the real-world data (Table 2). Hence, perhaps the most important lesson from this exercise is that the real-world data are quite heterogeneous and qualitatively different from simulated data with artificial cycles. The analyst should use extreme caution when applying a pre-trained model to new datasets.

Artificial Cycles with Noise. How do performances change if we add noise to the simulated data? Figure 6 shows examples of simulated cycles with noise. Tables 6, 7, and 8 report results when small, medium, and large white noises are added to the artificial cycles, respectively. We do not add noise to the first "white noise" simulation because it is already pure noise. Nevertheless, the three tables keep listing the same white-noise results as a reminder that some of the predictions are degenerate (i.e., full of false positives).

One might expect monotonically decreasing performances as we increase the noise level, which we operationalize as the standard deviation of the i.i.d. normal distribution. However, Table 6 shows small noise (standard deviation = 0.05) could actually help some of the existing methods that rely on asymmetry. Method 1 (trained in the WA data) now detects at least 10% of theoretical Edgeworth cycles, and Method 3 (also trained in the WA data) detects 93% of them, even though neither method could detect any *noiseless* Edgeworth cycles (Table 5). The reason is that the noise breaks the dominance of zero-price-change days in theoretical Edgeworth cycles. The performance of Method 4 also improves, but its predictions are mostly degenerate (i.e., it classifies 100% of pure white noise as cycles) anyway. The additional noise mechanically increases the number of big price increases, which triggers this method to flag more cycles.

By contrast, Methods 5 and 6 are hardly affected by small noise. These spectral models correctly dismiss noise as noise because white noise does not contain any systematic frequency component. Method 7 is unaffected, but that is because it typically finds cycles in either

Figure 6: Examples of Simulated Cycles with Noise



Note: These pictures show examples of simulated price series before we transform them to match the mean and standard deviation of each dataset. The horizontal axes represent calendar days.

100% or 0% of the cases due to its simple rule. The impact of noise on Methods 8–10 is mixed. Noise decreases the percentages of detected Edgeworth cycles but often increases those of reverse Edgeworth cycles and 3-day sine waves. Overall, a little bit of noise could be either good or bad, or have no effect, depending on the nature of simulated cycles and detection algorithms.

Table 7 shows medium-sized noise (standard deviation = 0.25) induces broadly similar patterns, albeit with some differences. For example, Method 1 stops working. Method 3 seems to perform reasonably well, but it now detects only 72% of Edgeworth cycles, instead of 93% with small noise. Moreover, it starts detecting cycles in other simulations by chance (in the top panel) even though its NMC criterion is supposed to capture only Edgeworth-

type asymmetry. Methods 5 and 6 still perform well (in the top and middle panel) but now detect only 63%–82% of the asymmetric cycles. Methods 8–10 are also negatively affected in most cases.

Finally, Table 8 reports the results under large noise (standard deviation = 0.5). Noise of any size is better than no noise for the use of Method 3 on theoretical Edgeworth cycles, but most of its predictions are now fooled by the presence of random shocks. Methods 5 and 6 are the only ones that continue to deliver valid classifications, albeit with the relatively low accuracy of 29%–39% for the asymmetric cycles.

Table 6: Simulated Data with Small Noise (Standard Deviation = 0.05)

Method	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
I. All Models Trained with Labeled Data from Western Australia										
White noise	0	0	36	100	25	19	100	0	0	0
Edgeworth	10	0	93	100	90	92	100	22	7	17
Reverse Edgeworth	0	0	0	96	90	91	100	0	8	0
Sine wave: 3 days	0	0	27	100	100	100	100	0	0	0
Sine wave: 7 days	0	0	34	87	100	100	100	72	2	53
Sine wave: 14 days	0	0	7	17	100	100	0	22	0	26
Sine wave: 21 days	4	0	0	0	100	100	0	2	0	0
Sine wave: 28 days	6	0	0	0	100	100	0	0	0	0
II. All Models Trained with Labeled Data from New South Wales										
White noise	100	0	80	100	22	15	100	0	3	0
Edgeworth	100	7	100	100	97	98	100	0	2	0
Reverse Edgeworth	100	0	43	57	98	99	100	0	1	0
Sine wave: 3 days	100	0	92	100	100	100	100	0	0	0
Sine wave: 7 days	100	0	99	30	100	100	100	0	27	0
Sine wave: 14 days	100	0	99	0	100	100	100	0	99	0
Sine wave: 21 days	100	0	99	0	100	100	100	78	99	100
Sine wave: 28 days	100	0	100	0	100	100	100	100	100	45
III. All Models Trained with Labeled Data from Germany										
White noise	0	24	17	100	0	0	100	85	4	100
Edgeworth	0	93	48	99	0	0	100	85	41	100
Reverse Edgeworth	0	0	0	100	0	0	100	96	5	43
Sine wave: 3 days	0	7	5	100	0	0	100	55	14	100
Sine wave: 7 days	0	1	5	100	0	4	100	38	0	100
Sine wave: 14 days	0	0	0	100	0	4	0	0	0	2
Sine wave: 21 days	0	0	0	92	0	4	0	0	0	0
Sine wave: 28 days	0	0	0	46	0	4	0	0	0	0

Note: The "white noise" simulations and results are the same as in Table 5 (with standard deviation = 1 in the original simulation). We list them here for reference—as a reminder that some of the predictions are degenerate. Each result (%) is based on 10,000 quarters of simulated data. See the text for details.

Table 7: Simulated Data with Medium Noise (Standard Deviation = 0.25)

Method	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
I. All Models Trained with Labeled Data from Western Australia										
White noise	0	0	36	100	25	19	100	0	0	0
Edgeworth	0	0	72	100	64	63	100	0	0	0
Reverse Edgeworth	0	0	10	100	64	63	100	0	1	0
Sine wave: 3 days	0	0	40	100	100	100	100	0	1	0
Sine wave: 7 days	0	0	40	99	100	100	100	56	3	27
Sine wave: 14 days	0	0	20	51	100	100	0	1	0	19
Sine wave: 21 days	0	0	11	34	100	100	0	0	0	0
Sine wave: 28 days	0	0	11	30	100	100	0	0	0	0
II. All Models Trained with Labeled Data from New South Wales										
White noise	100	0	80	100	22	15	100	0	3	0
Edgeworth	100	1	97	100	78	82	100	0	1	0
Reverse Edgeworth	100	0	50	100	77	82	100	0	0	0
Sine wave: 3 days	100	0	75	100	100	100	100	0	0	0
Sine wave: 7 days	100	0	90	91	100	100	100	0	29	0
Sine wave: 14 days	100	0	95	32	100	100	100	0	93	0
Sine wave: 21 days	100	0	95	19	100	100	100	0	99	0
Sine wave: 28 days	100	0	96	16	100	100	100	0	99	8
III. All Models Trained with Labeled Data from Germany										
White noise	0	24	17	100	0	0	100	85	4	100
Edgeworth	0	58	46	100	0	0	100	77	26	100
Reverse Edgeworth	0	5	3	100	0	0	100	90	6	99
Sine wave: 3 days	0	23	24	100	0	0	100	56	33	100
Sine wave: 7 days	0	12	16	100	0	4	100	61	0	100
Sine wave: 14 days	0	9	2	100	0	4	0	78	0	2
Sine wave: 21 days	0	7	1	100	0	4	0	13	0	0
Sine wave: 28 days	0	6	1	100	0	4	0	18	0	0

Note: The "white noise" simulations and results are the same as in Table 5 (with standard deviation = 1 in the original simulation). We list them here for reference—as a reminder that some of the predictions are degenerate. Each result (%) is based on 10,000 quarters of simulated data. See the text for details.

Table 8: Simulated Data with Large Noise (Standard Deviation = 0.5)

Method	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
I. All Models Trained with Labeled Data from Western Australia										
White noise	0	0	36	100	25	19	100	0	0	0
Edgeworth	0	0	44	100	34	29	100	0	0	0
Reverse Edgeworth	0	0	29	100	34	29	100	0	0	0
Sine wave: 3 days	0	0	40	100	100	95	100	0	3	0
Sine wave: 7 days	0	0	38	100	100	100	100	2	4	1
Sine wave: 14 days	0	0	27	97	98	100	63	0	0	0
Sine wave: 21 days	0	0	22	95	100	100	49	0	0	0
Sine wave: 28 days	0	0	22	93	100	100	48	0	0	0
II. All Models Trained with Labeled Data from New South Wales										
White noise	100	0	80	100	22	15	100	0	3	0
Edgeworth	100	0	87	100	39	35	100	0	2	0
Reverse Edgeworth	100	0	75	100	38	35	100	0	1	0
Sine wave: 3 days	100	0	73	100	100	100	100	0	0	0
Sine wave: 7 days	100	0	87	100	100	100	100	0	4	0
Sine wave: 14 days	100	0	88	93	100	100	100	0	48	0
Sine wave: 21 days	100	0	88	88	100	100	100	0	55	0
Sine wave: 28 days	100	0	89	87	100	100	100	0	62	0
III. All Models Trained with Labeled Data from Germany										
White noise	0	24	17	100	0	0	100	85	4	100
Edgeworth	0	31	22	100	0	0	100	83	11	100
Reverse Edgeworth	0	19	13	100	0	0	100	87	7	100
Sine wave: 3 days	0	27	25	100	0	0	100	62	47	100
Sine wave: 7 days	0	18	16	100	0	4	100	79	0	100
Sine wave: 14 days	0	17	7	100	0	4	44	93	0	69
Sine wave: 21 days	0	16	5	100	0	4	33	94	0	35
Sine wave: 28 days	0	15	4	100	0	4	30	96	0	17

Note: The "white noise" simulations and results are the same as in Table 5 (with standard deviation = 1 in the original simulation). We list them here for reference—as a reminder that some of the predictions are degenerate. Each result (%) is based on 10,000 quarters of simulated data. See the text for details.

Appendix B Additional Results

B.1 Combining Methods 1–4

This section investigates whether combining some or all of the existing methods leads to better performances. We construct 11 combinatorial methods based on Methods 1–4. Each combination comes in two specifications, AND and OR, depending on the logical operator combining its constituent methods. For example, the two variants of combination (7) in Table 9 are "Methods 1 AND 2 AND 3" and "Methods 1 OR 2 OR 3." The former detects cycles if all of Methods 1–3 do; the latter detects cycles if any of Methods 1–3 does.

Table 9: Accuracy (%) of Combinatorial Methods

Combination	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Constituent methods:											
1. PRNR	Yes	Yes	Yes	–	–	–	Yes	Yes	Yes	–	Yes
2. MIMD	Yes	–	–	Yes	Yes	–	Yes	Yes	–	Yes	Yes
3. NMC	–	Yes	–	Yes	–	Yes	Yes	–	Yes	Yes	Yes
4. MBPI	–	–	Yes	–	Yes	Yes	–	Yes	Yes	Yes	Yes
I. Western Australia (# manually labeled observations: 24,569)											
AND	90.34	90.88	90.89	91.18	91.74	89.58	90.34	90.33	90.89	91.18	90.33
OR	91.92	89.38	90.28	89.58	89.93	90.10	89.58	89.91	90.10	89.74	89.74
II. New South Wales (# manually labeled observations: 9,693)											
AND	81.18	78.74	81.91	78.62	81.86	81.55	81.18	82.08	81.91	81.86	82.08
OR	76.18	70.96	78.38	70.96	78.31	70.96	70.96	76.03	70.96	70.96	70.96
III. Germany (# manually labeled observations: 35,685)											
AND	60.44	60.44	60.44	60.49	60.84	60.48	60.44	60.44	60.44	60.49	60.44
OR	60.69	60.48	65.48	60.68	65.34	65.48	60.68	65.34	65.48	65.34	65.34

Note: "AND" and "OR" mean the constituent methods are combined with "and" and "or" operators, respectively. 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, whereas the latter subsample is used to evaluate the accuracy of predictions. All accuracy statistics are the medians from the 101 testing subsamples.

Table 9 shows that the performances of these combinatorial methods are similar to those of their constituent methods. The ranges of median accuracy results are 89%–92% in WA, 71%–82% in NSW, and 60%–65% in Germany, which are almost identical to those of individual Methods 1–4 in Table 2. Thus, combinations do not generate materially different predictions.

B.2 Variants of Methods 5–7

Table 10 reports the performances of the variants of Methods 5 (FT), 6 (LS), and 7 (CS). In Methods 5 and 6, the "max" and "HHI" variants are as explained in section 4.2 and Appendix A.1. 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 4.2. Its "integral" and "absolute value" variants are explained in Appendix A.1.

B.3 Data Requirement and Marginal Cost of Accuracty

Table 11 reports the means of accuracy (% correct) across 101 bootstrap sample splits that are underlying the visual summaries in Figure 2 in section 6. Table 12 shows the standard deviations of accuracy are usually less than 1 percentage point when more than 1% of the sample is used for training.

Figure 7 and Panel (C) of Table 11 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 5 are encouraging in that only a few hundred labeled observations are necessary to reach approximately optimal accuracy levels.

Table 10: 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 sections A.1 and B.2 for the definition of each method.

Table 11: 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

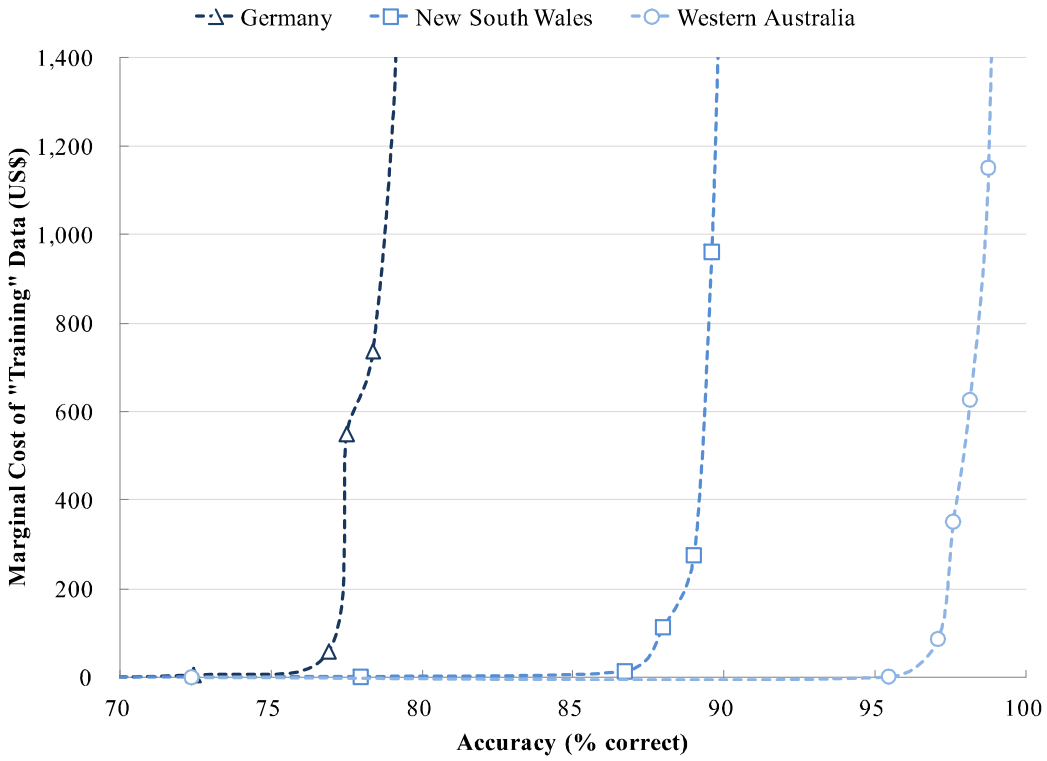
Note: The numbers in panel (A) indicate accuracy in the 20% testing subsample. Some or all of the remaining 80% subsample are used as a training subsample to optimize the parameters of each model. We randomly split the sample 101 times, tune the parameters as many times, and report their median performances. 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 3).

Table 12: 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 7: Marginal Costs of Accuracy (E-LSTM)



Note: These marginal-cost curves are based on the numbers reported in Table 11 (shown with markers) and a splines-based interpolation (dashed lines).

B.4 Using Only "Cleaner" Subsamples

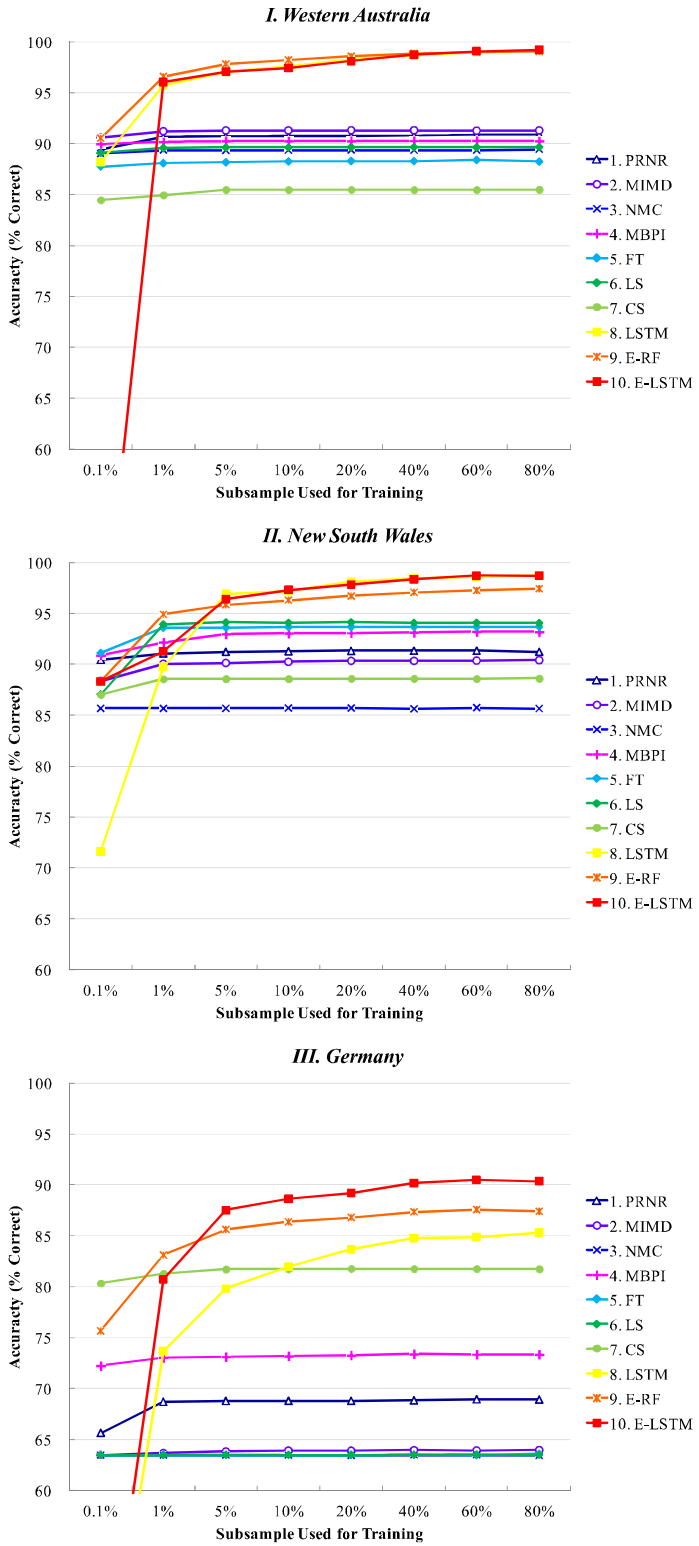
Table 13 and Figure 8 report alternative results based only on subsamples that are either unanimously labeled as "cycling" by all RAs or not labeled as "cycling" by any RA, thereby ignoring ambiguous observations. Accuracy is higher overall, but relative performance rankings remain similar to the baseline results.

Table 13: 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 8: 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.

B.5 Ambiguous Cycles with Mixed Labels

This section investigates the extent to which algorithmic classifications reflect the ambiguity in human labels. The examples in Figure 1 are clear cycles or non-cycles that our RAs unanimously labeled as such. However, our data also contain many ambiguous cases in which some or all RAs chose "maybe cycling" instead of "cycling" or "not cycling." We codify only those station-quarters with unanimous "cycling" labels as $cycle_{i,t} = 1$, but readers might be curious how the algorithms respond to ambiguous cases in the data.

The first two columns of Table 14 list ambiguous combinations of manual labels and the count of station-quarter observations in each category, respectively. All other columns report the percentages of observations that are classified as cycles ($\widehat{cycle}_{i,t} = 1$) by the 10 algorithms. Technically speaking, the combination of $\widehat{cycle}_{i,t} = 1$ and $cycle_{i,t} = 0$ constitutes a "false positive" and is treated as a prediction error in our main analysis. Nevertheless, for the purpose of this section, we can also interpret the false-positive rate as a measure of how these methods reflect the underlying ambiguity in data and human judgment.

Table 14: Percentage of Ambiguous Cases Classified as Cycles

Label composition	Count	(1) PRNR	(2) MIMD	(3) NMC	(4) MBPI	(5) FT	(6) LS	(7) CS	(8) LSTM	(9) E-RF	(10) E-LSTM
II. New South Wales (# manually labeled observations: 9,693)											
(*yes*, *yes*, *maybe*)	843	83.7	100.0	73.2	64.5	75.6	76.2	99.1	12.9	52.6	54.9
(*yes*, *maybe*, *maybe*)	412	78.4	100.0	65.3	56.1	65.5	63.3	97.8	7.5	23.3	24.3
(*yes*, *maybe*, *no*)	238	63.9	100.0	60.1	41.6	46.2	44.1	95.0	3.4	6.3	10.9
(*maybe*, *maybe*, *maybe*)	118	65.3	100.0	75.4	45.8	55.9	49.2	89.0	4.2	8.5	9.3
(*maybe*, *maybe*, *no*)	218	55.0	100.0	71.1	22.9	30.3	26.6	80.3	0.5	0.5	2.3
(*maybe*, *no*, *no*)	415	38.8	100.0	48.2	14.5	18.3	15.2	77.1	0.0	0.2	0.2
Mix of *yes* & *no*	172	71.5	100.0	57.0	36.0	47.1	43.0	90.1	6.4	9.3	11.0
III. Germany (# manually labeled observations: 35,685)											
(*yes*, *yes*, *maybe*)	6,986	0.0	0.2	2.2	22.6	0.0	0.0	37.5	5.5	29.5	40.9
(*yes*, *maybe*, *maybe*)	5,511	0.0	0.2	1.7	16.7	0.0	0.0	22.9	1.9	13.8	15.3
(*yes*, *maybe*, *no*)	748	0.0	0.1	1.9	12.8	0.0	0.0	15.9	1.6	7.6	9.8
(*maybe*, *maybe*, *maybe*)	4,435	0.0	0.1	1.4	12.1	0.0	0.0	16.7	1.0	6.6	6.0
(*maybe*, *maybe*, *no*)	2,161	0.0	0.0	2.0	9.4	0.0	0.0	10.9	0.5	3.3	3.3
(*maybe*, *no*, *no*)	1,013	0.0	0.0	3.8	4.3	0.0	0.0	3.7	0.1	0.7	0.4
Mix of *yes* & *no*	402	0.0	0.2	2.7	21.6	0.0	0.0	39.6	7.5	32.8	41.0

Note: "Yes," "maybe," and "no" correspond to "cycling," "maybe cycling," and "not cycling" in the original human labels, respectively. The count of observations does not add up to the total count of manually labeled observations because "unambiguous" cases (i.e., unanimous "yes" or "no") are excluded.

The results suggest many methods respond to ambiguous cases in an intuitive manner.

For example, Methods 9 (E-RF) and 10 (E-LSTM) respectively classify 53% and 55% of the $(yes, yes, maybe)$ cases in NSW as cycles, which seems understandable. Likewise, they classify 23% and 24% of the $(yes, maybe, maybe)$ cases in NSW as cycles. The fraction of such "false positives" decreases as the label composition becomes more negative (i.e., closer to unanimous "no"). Similar patterns emerge elsewhere as long as the algorithmic predictions are non-degenerate. Thus, not all "prediction errors" are necessarily bad, as they could reflect subtle differences in the underlying data and manual labels.

B.6. Heterogeneity among Human Labelers

This section investigates the extents to which human labelers are heterogeneous and how such heterogeneity could affect the cycle-detection algorithms.

Are some labelers "stricter" or "looser" than others in recognizing cycles? The answer is "yes." Table 15 shows the most stringent RA labels 75% of the NSW as cycling, whereas the least stringent RA labels 87% as cycling.[58] Likewise, 55%, 59%, and 62% of the German data are flagged as cycling by individual RAs, respectively.[59]

Table 15: Heterogeneity among Human Labelers

Dataset	New South Wales			Germany		
Labeler ID#	1	2	3	1	2	3
Number of labels						
"Cycling"	7,295	7,482	8,428	19,711	21,084	22,112
"Maybe cycling"	1,114	1,158	838	12,072	12,905	12,141
"Not cycling"	1,284	1,053	427	3,902	1,696	1,432
Fraction of "cycling"	75.3%	77.2%	87.0%	55.2%	59.1%	62.0%
Profit margins (cent)						
"Cycling": mean	11.93	11.95	11.65	98.18	98.22	98.24
"Cycling": std. dev.	5.55	5.62	5.79	3.57	3.56	3.64
Others: mean	10.85	10.70	11.76	98.81	98.82	98.82
Others: std. dev.	7.27	7.21	7.49	4.61	4.71	4.69
Difference in means	1.08	1.25	−0.11	−0.63	−0.60	−0.58
p value	$< .001$	$< .001$	.617	$< .001$	$< .001$	$< .001$

Note: The total number of manually labeled observations may not be identical to the ones in Table 1 because this table omits observations with more than three labels.

Do the profit-margin statistics (the average difference between "cycling" and other observations) vary across individual labelers? The answer is "yes" again. The mean difference

[58] We do not study labeler heterogeneity in the WA data, where each station-quarter observation is labeled only once based on the consensus of two RAs.

[59] The German data were labeled by six RAs in total, but most observations have only three labels. We treat "label 1," "label 2," and "label 3" in our anonymized dataset as individual labelers.

is 1.24 cents in NSW when we use the three RAs' consensus (in our baseline results in Table 3); it is 1.08, 1.25, and −0.11 cents according to individual labelers 1, 2, and 3, respectively. Labeler 3 (the least stringent RA) turns out to be an outlier in this respect. By contrast, the impact of labeler heterogeneity on margin gaps is much less pronounced in the German data. The mean difference is −0.47 cents based on the consensus of RAs (Table 3), whereas the classifications by individual RAs lead to −0.63, −0.60, and −0.58 cents of margin gaps, respectively.

The heterogeneity across human labelers translates into heterogeneity in the performance of the algorithms. Table 16 summarizes the results of training each method based on each individual RA's labels alone. Its upper half shows the accuracy (%) in terms of correctly matching human labels. The general level of accuracy is high and comparable to the consensus-based baseline results (Table 2), but some methods seem to work "better" with individual labels in NSW than with the consensus version. Curiously, the accuracy performance systematically improves as algorithms try to replicate less stringent labelers, with the exception of Methods 7–10 in Germany. This pattern suggests less stringent labelers tend to follow some simple decision rules that are easily captured by the automatic methods.

The algorithmic versions of the margin-gap statistics generally agree with those based on their respective human-label targets, but there are exceptions. First, Methods 3 and 6 produce degenerate classifications, which precludes the calculation of differences in the first place. Second, Methods 2, 4, and 7 tend to produce the mean differences that are visibly different from the manual versions in NSW. Curiously, Methods 5 and 6 lead to positive gaps (0.19 and 0.29) for labeler 3 in NSW, even though this RA's labels generate a negative gap (−0.11).[60] Third, Methods 1, 2, and 5 also lead to visibly different statistics in Germany.

In summary, we find that different training data (labels) lead to different automatic-classification results. This finding is trivial and mechanical because the algorithms are designed to replicate manual labels. Nevertheless, it also highlights the importance of not relying too much on any single labeler. Human recognition is heterogeneous, and machines may replicate such heterogeneity. Thus, the real question is not so much about the methods themselves as about establishing a commonsensical "ground truth" by humans.

[60] This discrepancy is not entirely surprising because the p value of the negative gap is high (0.617). That is, the difference is not statistically significant at any conventional level.

Table 16: When Algorithms Are Trained by Individual Labelers

Dataset	New South Wales			Germany		
Labeler ID#	1	2	3	1	2	3
Accuracy (% correct)						
Method 0: Manual	100.00	100.00	100.00	100.00	100.00	100.00
Method 1: PRNR	81.63	82.94	90.79	60.28	62.95	64.49
Method 2: MIMD	80.64	82.46	89.97	56.30	59.46	62.00
Method 3: NMC	75.26	77.19	86.95	55.24	59.08	61.96
Method 4: MBPI	83.02	84.51	92.39	64.62	65.92	66.17
Method 5: FT	83.98	85.84	92.22	55.24	59.09	61.97
Method 6: LS	84.49	86.01	92.19	55.24	59.08	61.96
Method 7: CS	78.03	79.81	89.10	71.48	72.66	70.12
Method 8: LSTM	89.71	91.99	96.06	73.95	74.30	72.69
Method 9: E-RF	97.52	97.69	98.90	95.25	95.22	94.53
Method 10: E-LSTM	90.50	92.01	95.88	78.08	78.19	74.91
Difference in mean profit margins (cent)						
Method 0: Manual	1.08	1.25	−0.11	−0.63	−0.60	−0.58
Method 1: PRNR	0.42	0.40	−0.40	−0.78	−0.88	−1.49
Method 2: MIMD	3.93	3.37	−1.40	−0.63	−0.92	−3.96
Method 3: NMC	–	–	–	–	–	–
Method 4: MBPI	2.56	0.76	−1.03	−0.25	−0.75	−1.23
Method 5: FT	1.25	1.10	0.19	2.52	2.52	2.52
Method 6: LS	1.25	1.03	0.29	–	–	–
Method 7: CS	−1.90	−1.90	−1.90	−0.68	−0.74	−0.90
Method 8: LSTM	1.48	1.80	−0.18	−1.01	−0.88	−1.01
Method 9: E-RF	1.05	1.26	−0.09	−0.69	−0.64	−0.62
Method 10: E-LSTM	1.53	1.58	−0.17	−0.70	−0.88	−0.86

Note: No profit-margin-gap results are shown when classification results are degenerate.

Appendix C Supplementary Plots for Section 7.2

C.1. Why Existing Methods Work in Australia But Fail in Germany Figure 9 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).

C.2. Why Margins Correlate Positively with Cycles in Australia But Negatively in Germany The scatter plots of Figure 10 show the means and the standard deviations of margins are positively correlated in all datasets. As the histograms of Figure 11 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.

C.3. Why Do Existing Methods Find "Positive Correlations"? Figure 12 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 10 shows, however, high volatility is a poor predictor of true cyclicality (based on manual classification) in the German data.

Figure 9: Histograms of Median Daily Change

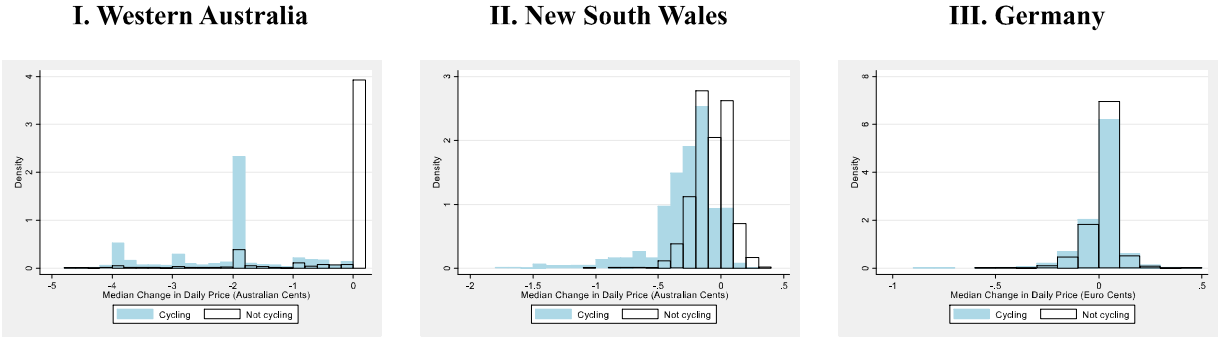


Figure 10: Scatter Plots of Mean and Standard Deviation

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

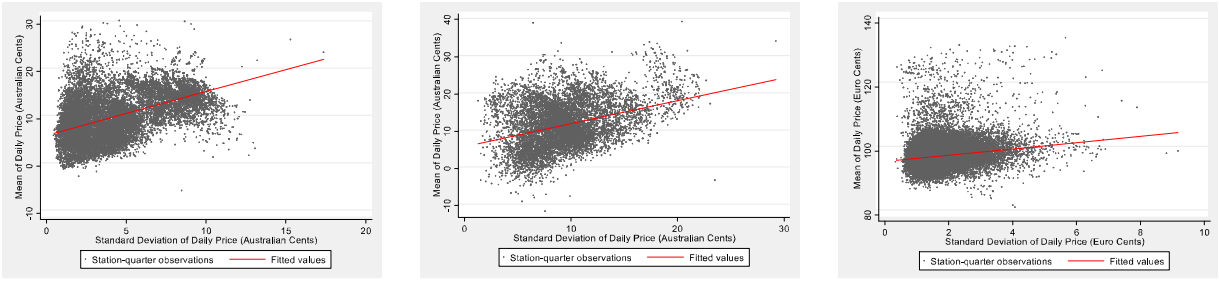


Figure 11: Histograms of Standard Deviation

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

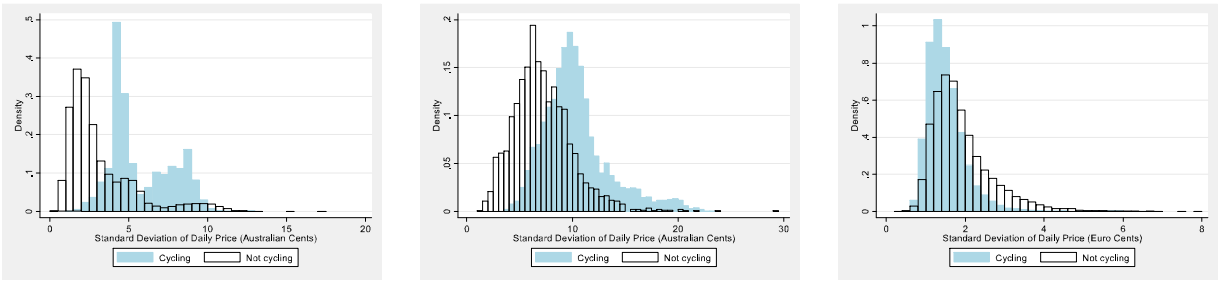
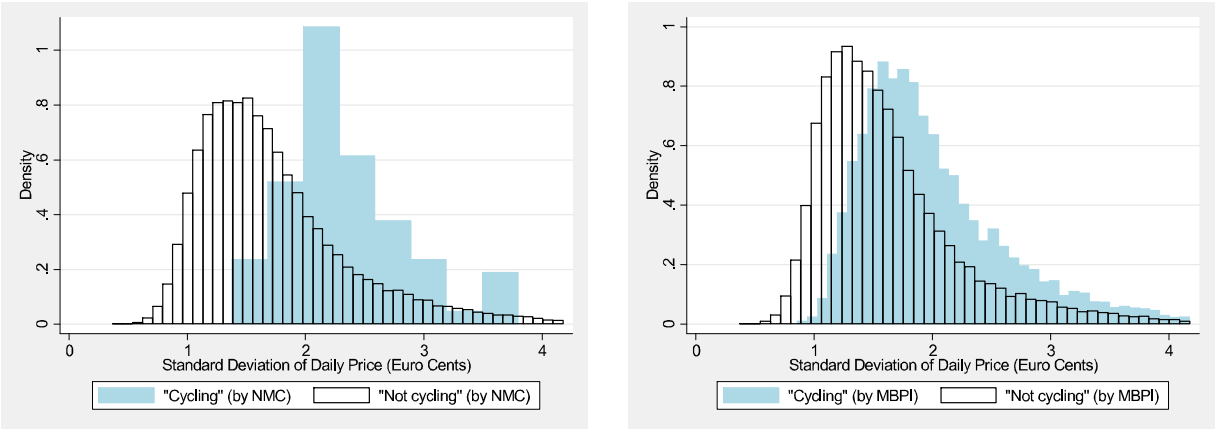


Figure 12: Asymmetry-based Definitions Pick Up Volatility

Method 3 (Negative Median Change) **Method 4 (Many Big Price Increases)**



Appendix D Additional Data Exploration

This section reports additional findings from the exploratory data analysis in section 7.3.

D.1 Additional Time Series Patterns: Macroeconomic Shocks

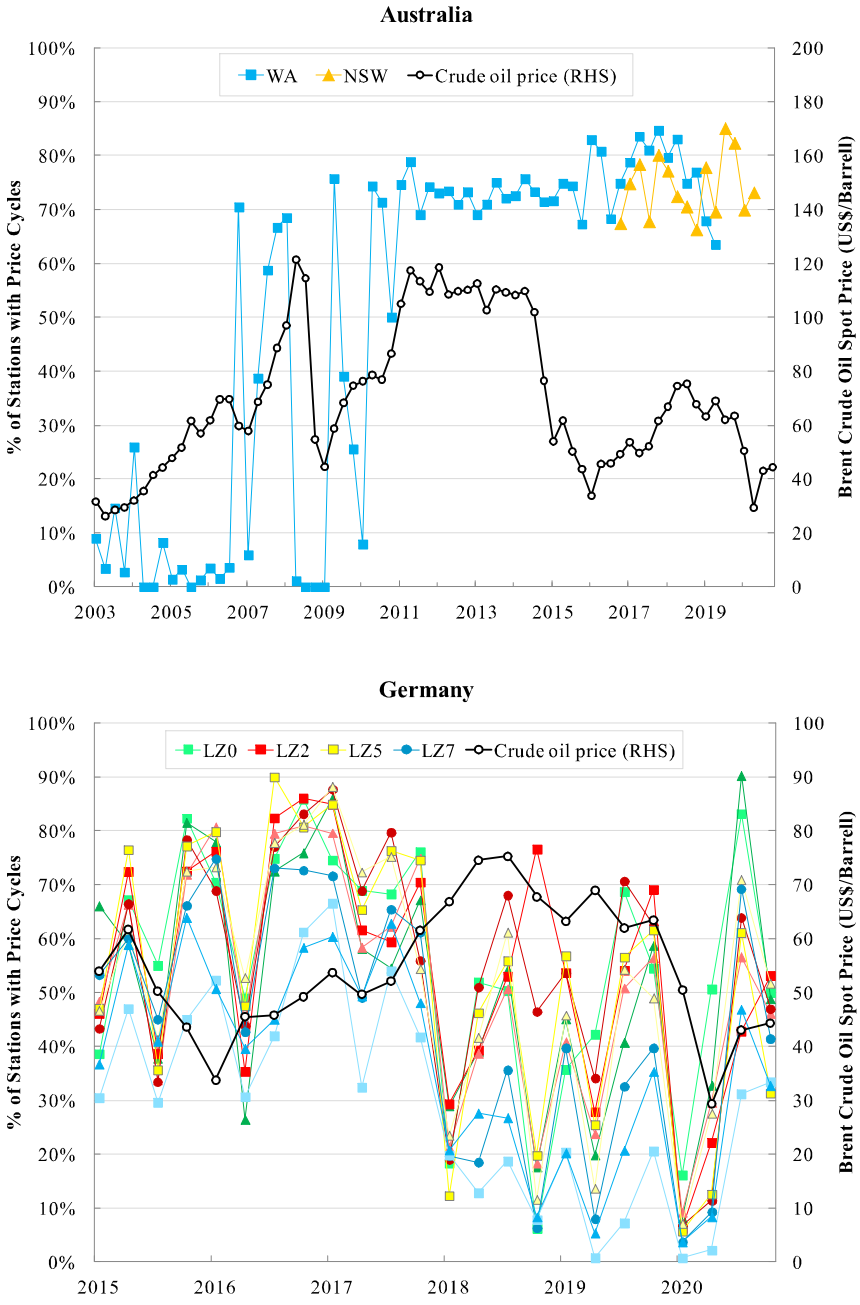
The relationship between macroeconomic shocks and price cycles seems complex. On the one hand, the "great recession" of 2008 roughly corresponds to the period of almost no cycles in WA, in a stark contrast with the adjacent periods in which cycles are prevalent. On the other hand, the onset of the covid-19 pandemic in 2020:Q1 does not seem to trigger any major change in the prevalence of cycles in NSW. Likewise, in Germany, the big drop in 2020:Q1 seems to correspond to the beginning of the pandemic, but this episode is hardly unique. There are several other big drops (e.g., 2015:Q3, 2016:Q2, and 2018:Q1) that are not clearly associated with business cycles.

Figure 13 overlays the fraction-of-cycling-stations data with crude oil price. Here again, the relationship is complicated. The prevalence of price cycles in WA does seem to correlate with oil price at first glance. The first "boom" of cycles in 2007 coincides with a large increase in oil price from approximately \$50 to \$100. The second boom from around 2010 also coincides with another increase in oil price. However, their ups and downs in 2008–2009 are not synchronized, and some of the biggest drops in oil price in recent history (in 2014–2015) fail to have any impact on the prevalence of price cycles.[61] In Germany, it is also possible to see *some* correlations but not decisive ones.[62] Thus, we tentatively conclude that serious investigations into the relationships between price cycles, input costs, and business cycles would require a deeper understanding of the mechanism that generates price cycles.

[61] Oil price's correlation with the fraction of cycling stations is 0.37 in WA and 0.06 in NSW.

[62] In fact the fractions of cycling stations in all of the 10 German regions are negatively correlated with crude oil price (between -0.05 and -0.40).

Figure 13: Crude Oil Price and Prevalence of Price Cycles

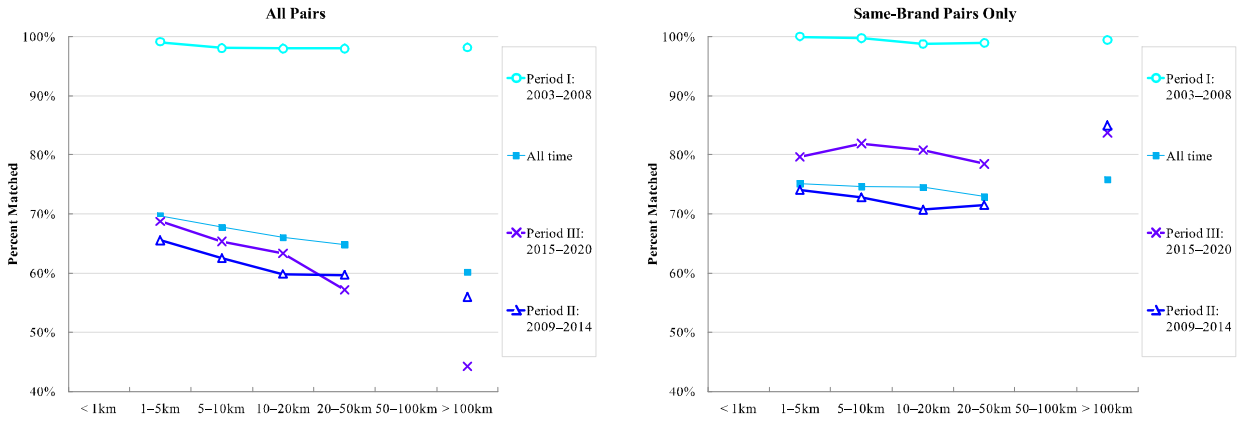


Note: Crude oil price is quarterly average of the monthly data on "Europe Brent Spot Price FOB" from Thomson Reuters downloaded from the US Energy Information Administration. The legend shows only four of the 10 LZs due to space limit.

D.2 Additional Spatial Patterns: Cycle Correlation and Distance in WA by Subperiod

WA's sample period is considerably longer than NSW's and Germany's and spans qualitatively different subperiods. In Figure 14, we further investigate whether patterns change over time in WA by showing "correlations" for each of the three six-year subperiods: 2003–2008, 2009–2014, and 2015–2020. The first subperiod (Period I) ends before clear weekly cycles emerged. Hence, the extremely high "correlation" is largely an artifact of many station-pair-quarter observations whose cycle statuses trivially match (i.e., many observations do not show cycles at all). By contrast, Period II roughly correspond to the era in which weekly cycles became prominent, and Period III is when crude oil price became volatile again but gasoline price cycles remained prevalent (see Appendix D.1). This "decomposition shows that Period I is an outlier for our purposes and that the all-time averages are closer to Periods II and III.[63]

Figure 14: Cycle Correlation in WA by Subperiod



Note: The WA graphs do not show markers or lines in two distance bins (less than 1km and 50–100km) because the WA data have relatively few pairs in these bins, which we grouped in the adjacent bins for the purpose of visualization. The number of valid observations (i.e., station pairs with at least 12 quarters of data in common) underlying the all-time statistics is larger than the sum of valid observations in Periods I–III.

[63] The all-time averages are closer to Periods II and III partly because more valid station-pair observations are available in later years. Note the number of valid observations (i.e., station pairs with at least 12 quarters of data in common) underlying the all-time statistics is larger than the sum of valid observations in Periods I–III, which is why the all-time percentage is lower than those of Periods I, II, and III in the right panel of Figure 14 (in the largest-distance bin).

Appendix E Potential Biases in Adversarial Circumstances

This section discusses the extent to which adversarial circumstances (e.g., in antitrust cases) could affect (i) the manual labeling of data, (ii) the choice of automatic detection methods, and (iii) the descriptive findings about cycles and profit margins.

To make our discussions concrete, we assume two adversaries, Plaintiff ("P") and Defendant ("D"), dispute whether price cycles are anti-competitive. P's objective is to prove that (A) price cycles exist in the data and that (B) price cycles positively correlate with profit margins. D's objective is to prove that (A') price cycles do not exist in the data and that (B') even if they did, they are either negatively correlated or simply uncorrelated with profit margins. Because a statistically significant *negative* correlation could potentially invite another scrutiny (i.e., "Does the absence of cycles indicate lack of competition?"), avoiding any systematic pattern would be D's best defense.

Let us first consider how P and D will approach task (i), the manual labeling of data. As Appendix C.4 illustrates, human labelers make heterogeneous judgments even under the direction of a single supervisor and working in a team with regular meetings to coordinate their classification criteria (see section 3.2). Hence, P and D can select their respective teams of experts based on their tendencies in manual labeling. More direct maneuvering is possible as well. P can instruct its experts to label relatively high-margin observations as "cycles" and low-margin ones as "non-cycles." Likewise, D would incentivize its own experts to randomize their labels, so that the labeled data become too noisy to show any systematic patterns. Therefore, an obvious conclusion for task (i) is that it must be performed by an independent, disinterested party and that the same "ground truth" (labeled dataset) must be used by both P and D. Otherwise, generating arbitrary labels will always be possible.

Once the "ground truth" is established, (ii) the choice of automatic methods is relatively straightforward because it is simply a matter of training and choosing a method that achieves the highest accuracy in replicating the ground truth without degenerate predictions. Accuracy is a good criterion as long as the ground truth is shared by P and D. Our results (Tables 2 and 3) show that "wrong" results are by-products of inaccurate classifications. Among the 10 methods that we evaluated, Method 10 would be the most desirable in this respect. Nevertheless, preparing several other methods and comparing their outputs would be advisable as a sensitivity analysis.[64]

[64] If similar methods—or models with similar levels of complexity and accuracy—generate radically differ-

Task (iii)—the calculation of descriptive statistics about cycles and margins—is straightforward as well. After tasks (i) and (ii) are complete, the scope for manipulating statistical finding is limited. The "only" remaining question is whether and how to control for additional observable characteristics of shops/firms and time periods. In datasets with many observed characteristics (and potential heterogeneity across units observations), determining the appropriate level of aggregation and comparison would require nontrivial efforts. But these problems are common to any empirical analysis, well-understood by seasoned economists, and not unique to the detection of price cycles. Hence, we do not discuss them here.

In summary, adversarial biases are most likely to be an issue in the manual labeling stage rather than in the later, more technical parts of the analysis. Thus, the selection of human labelers should be treated in the same spirit as in the selection of jury in trials.

ent patterns, that might be a symptom of issues in data cleaning, coding errors, or other technical problems in the estimation/training of the models. Additional insights could be gained by comparing not only the accuracy levels but also the composition of prediction errors (false negatives vs. false positives).

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] Bundeskartellamt. 2011. *Fuel Sector Inquiry*. Final Report in accordance with §32e GWB - May 2011 - Summary.

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

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

[6] 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.

[7] 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.

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

[9] Chandra, Ambarish, and Mariano Tappata. 2011. "Consumer search and dynamic price dispersion: an application to gasoline markets," *RAND Journal of Economics*, 42 (4): 681–704.

[10] 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).

[11] 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.

[12] 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.

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

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

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

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

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

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

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

[20] 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.

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

[22] Haucap, Justus, Ulrich Heimeshoff, and Manuel Siekmann. 2017. "Fuel Prices and Station Heterogeneity on Retail Gasoline Markets," *Energy Journal*, 38 (6): 81–103.

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

[24] Holt, Timothy, Mitsuru Igami, and Simon Scheidegger. 2023. "Replication package for: Detecting Edgeworth Cycles (Version 2)." URL: https://dx.doi.org/10.5281/zenodo.10126406.

[25] Ivaldi, Marc, Bruno Jullien, Patrick Rey, Paul Seabright, and Jean Tirole. 2003. "The Economics of Tacit Collusion." Final Report for DG Competition, European Commission.

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

[27] 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.

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

[29] 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.

[30] Linder, Melissa. 2018. "Price cycles in the German retail gasoline market - Competition or collusion?" *Economics Bulletin*, 38 (1): 593–602.

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

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

[33] 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.

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

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

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

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

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

[39] 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.

[40] 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.

[41] Siekmann, Manuel. 2017. "Characteristics, causes, and price effects: Empirical evidence of intraday Edgeworth cycles." DICE Discussion Paper, No. 252.

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

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

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

[45] 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.