Detecting Edgeworth Cycles

Timothy Holt Università della Svizzera Italiana
Mitsuru Igami Yale University
Simon Scheidegger University of Lausanne

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 with data from Western Australia and New South Wales, but only a few can detect the nuanced cycles in Germany. Results suggest that 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.

1. Introduction

Retail gasoline prices are known to follow cyclical patterns in many countries (see, for example, 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 anticompetitive business practices. The occasional discovery of price-fixing cases supports this view (see, for example, Clark and Houde 2014; Foros and Steen 2013; Wang 2008).\(^1\)

Igami is also affiliated with the University of British Columbia. 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 (Alan Chiang, Jordan Mazza, Bruno Moscarini, Clara Penteado, Yue Qi, Alexis Teh, Janie Wu, and Eileen Yang) for excellent work. We also thank Roxana Mihet and seminar and conference participants at the Dusseldorf Institute for Competition Economics, Brown University, Cornell University, Yale University, the 2022 International Industrial Organization Conference, and the Econometric Society’s 2022 North American Summer Meeting for comments. This work was supported by the Swiss National Science Foundation under project ID “New Methods for Asset Pricing with Frictions.”

\(^1\) 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
These asymmetric movements are called Edgeworth cycles and have been studied extensively. 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, Yun, and Taylor (2013), and Noel (2015) show that prices and margins are lower in markets with asymmetric price cycles. Given the diversity of countries and regions in these studies (Australia, Canada, the United States, 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 data sets. 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, developing scalable detection methods represents an important practical challenge for economists and policy makers.

This paper proposes a systematic approach to detecting Edgeworth cycles. We formalize four existing methods as simple parametric models: (1) the positive runs versus negative runs (PRNR) method of Castanias and Johnson (1993), (2) the mean increase versus mean decrease (MIMD) method of Eckert (2002), (3) the negative median change (NMC) method of Lewis (2009), and (4) the many big price increases (MBPI) method of Byrne and de Roos (2019). We then pro-

---

2 Maskin and Tirole (1988) coined the term after the hypothetical example of Edgeworth (1925). It has been a popular topic for empirical research since Castanias and Johnson (1993). Its theoretical background is explained 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 (https://www.accc.gov.au/consumers/petrol-diesel-lpg/about-fuel-prices). The Bundeskartellamt does the same in Germany.

4 Systematic methods to detect price cycles are also useful for researchers who do not want to study cycles. Chandra and Tappata (2011, p. 697 n. 46) examine the role of consumer search in generating temporal dispersion in US retail gasoline prices but could not completely reject Edgeworth cycles as an alternative explanation because they did not have a scalable method to prove the absence of cycles in their large data set of more than 25,000 stations. Our procedure would have allowed them to provide more concrete evidence.
pose six new methods based on spectral analysis and nonparametric/machine-learning techniques: (5) Fourier transform (FT), (6) the Lomb-Scargle periodogram (LS), (7) cubic splines (CS), (8) long short-term memory (LSTM), (9) an ensemble (aggregation) of methods 1–7 within a random-forests framework (E-RF), and (10) an ensemble of methods 1–8 within an extended LSTM (E-LSTM).

To evaluate the performance of each method, we collect data on retail and wholesale gasoline prices in Germany and two regions of Australia: Western Australia and New South Wales. These data sets cover the universe of gasoline stations in these geographical areas, record each station’s retail price at a daily (or higher) frequency, and are made publicly available by legal mandates (on the Australian data, see Byrne, Nah, and Xue [2018]; on the German data, see, among others, Haucap, Heimeshoff, and Siekmann [2017]; Martin [2018]; Assad et al. [2024]). 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 (that is, retail minus wholesale prices) of gasoline stations and group them into calendar quarters, so that a station-quarter (that is, a set of 90 consecutive days of retail-margin observations for each station) becomes the effective unit of observation. Eight research assistants (RAs) manually classify each station-quarter as either “cycling,” “maybe cycling,” or “not cycling.” A binary indicator variable is used that equals one if an observation is labeled cycling by all of the RAs (the majority of observations as labeled by three RAs) and zero otherwise as a conservative target for automatic cycle detection.\textsuperscript{5} 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 mathematical definitions and can easily be checked at a later stage. Hence, we prioritize the detection of cyclicality, which alleviates the cognitive burden on RAs.

Figure 1 shows examples of cycles and noncycles in station-quarter observations. The vertical axes measure retail gasoline prices in Australian cents (Figures 1A and 1B) and euro cents (Figure 1C). The horizontal axes represent calendar days.

At this point, one might wonder whether human recognition of cycles is an appropriate benchmark. We regard it as the best feasible option given the lack of clear theoretical definitions. 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’ visual inspection 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 peren-

\textsuperscript{5} Sections OB4–OB6 in Online Appendix OB show results under alternative criteria.
A 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.

Three sets of results are reported. First, when applied to the Western Australia and New South Wales data sets, most of the methods—both existing and new—achieve high levels of accuracy near or above 90 percent and 80 percent, 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 percent of the sample is unanimously labeled cycling by three RAs (see Figure 1C for an example). 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 called the “price parachute” (or Preis Fallschirm) phenomena and are considered to be part of common pricing strategies among practitioners. The Bundeskartellamt (2011) confirms the existence of both weekly and daily cycles. Methods 7–10 attain 71–80 percent accuracy even in this challenging environment.

Second, the cost-effectiveness of each method is assessed by using only .1, 1, 5, 10, . . . , 80 percent of the manually labeled subsamples as training data. Results suggest that 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 requirements are 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 on 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 data set for 2014–20 would require 4,800 RA hours, or US$64,800. Thus, this approach is economical and suitable for researchers and governments with limited resources.

Third, whether and how gasoline stations’ markups are correlated with the presence of cycles is investigated. In Western Australia and New South Wales, the average margins in (manually classified) cycling station-quarters are statistically significantly higher than in noncycling ones. The relationship is reversed in Germany, where the margins in cycling observations are lower than in noncycling 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 (that is, positive correlations) in Western Australia, but some of them fail in New South Wales. Furthermore, methods 1–6 either fail to detect cycles or lead to false conclusions in Germany (that is, find statistically significant positive correlations). This finding emerges under both definitions of Edgeworth cycles based only on cyclicality and those that include asymmetry as well. 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. The remainder of this section clarifies the related literature, our contributions, and the replication package. Sections 2–4 explain the theoretical background, data, and methods, respectively. Sections 5–7 report the main findings and discuss their economic and policy implications. Section 8 summarizes the practical recommendations for cycle detection. Section 9 concludes.

Related Literature. This work is closely connected to the Edgeworth-cycle literature and cites many related works throughout the paper. This introductory section provides the overall literature context, Section 2 covers the theoretical background, Section 3 cites data sources and several papers that use the German data, Section 4.1 acknowledges the proponents of the existing methods, and Section 4.2 suggests helpful readings for the new methods proposed.

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

Replication Package. To lower the entry barriers for empirical economists considering the use of advanced machine-learning tools, we have made the com-
puter code (in Python) and the data set available as a replication package.\textsuperscript{6} Detailed documentation for nontechnical users is included.

2. Theoretical Background

The primary goal of this paper is empirical. Some conceptual anchoring will nevertheless clarify the target of measurement.

2.1. What Are Edgeworth Cycles?

Maskin and Tirole (1988, pp. 571–72) offer the following 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.” 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 the original conjecture by Edgeworth (1925). His writing focuses on the indeterminacy of static equilibria in a price-setting game between capacity-constrained duopolists. Even though he mentions a price path that resembles the description by Maskin and Tirole (1988) 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” (Edgeworth 1925, p. 118). Thus, Edgeworth’s original theory features not so much cyclicality as “perpetual motion” (p. 121).

Nevertheless, we have chosen to focus on cyclicality in this paper. Theoretically, the equilibrium strategies of Maskin and Tirole (1988, p. 587, 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 formalization of Maskin and Tirole (1988) and the subsequent empirical literature (see methods 1–4 in Section 4.1).

Property 3: Stochasticity. In the Edgeworth-cycle MPE of Maskin and Tirole (1988), large 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

\textsuperscript{6}The replication package and Online Appendix are available on the Journal’s website.
Detecting Edgeworth Cycles

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.\textsuperscript{7} We do not impose stochastic frequencies as a necessary condition in our empirical procedures, but some methods are designed to accommodate cycles with varying frequencies (see 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.\textsuperscript{8} 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 location, 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 data sets. Consequently, we do not impose any geographical boundaries a priori and simply analyze data at the individual station level.\textsuperscript{9} The idea is that once the station-level characterization is successfully completed, one can always compare cyclicity 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, the narrative by Edgeworth (1925) lacks any hint of cooperative actions or intentions. On the other hand, Maskin and Tirole (1988, p. 592) 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.” In the more recent literature how-

\textsuperscript{7} This theoretical property seems largely overlooked in the empirical literature, presumably because the first two properties make the phenomena sufficiently interesting and policy relevant.

\textsuperscript{8} Alternative explanations include consumers with heterogeneous search costs, intertemporal price discrimination, and dynamic pricing algorithms (broadly defined as any pricing strategy and its implementation that try to exploit consumer heterogeneity and time-varying price elasticity of demand).

\textsuperscript{9} 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 changing prices, it might not be possible to observe clear cycles in 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, the Edgeworth-cycle Markov perfect equilibria of Maskin and Tirole (1988) require a fine grid with sufficiently small intervals (denoted $k$ in their model). Therefore, we believe the risk of missing aggregate cycles is low.
ever the term “tacit collusion” is usually associated with collusive equilibria in repeated-games models. 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 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 documents 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 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 Western Australia, New South Wales, and Germany. We combine them with wholesale-price data, on the basis of the region of each station (Australia) or the location of the nearest refinery (Germany). Station-level daily profit margins are computed by subtracting the relevant wholesale price from the retail price,

\[ p_{i,d} = p_{i,d}^R - p_{i,d}^W, \]

where \( p_{i,d}^R \) and \( p_{i,d}^W \) are retail and wholesale prices at station \( i \) on day \( d \), and markup measure \( p_{i,d} \) is henceforth referred to as the price. Daily prices are organized by calendar quarter, so a station-quarter (that is, a sequence of daily prices over 90 days for each station) becomes the unit of observation for cycle detection.

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 et al. (2003).
Detecting Edgeworth Cycles

3.1. Data Sources and Preparation

Retail Prices. Three data sets of retail gasoline prices that are publicly available and of high quality are used. FuelWatch and FuelCheck are legislated retail-fuel-price platforms operated by the state governments of Western Australia and New South Wales, respectively. Their websites display real-time information on gas prices, and the complete data sets can be downloaded.\textsuperscript{11} The Market Transparency Unit for Fuels of the Bundeskartellamt publishes similar data for every German gas station in minute intervals.\textsuperscript{12}

Sampling Frequencies. The raw data from Western Australia contain daily retail prices for each station, which is the most granular level in this region, because its law mandates that each station must commit to a fixed price level for 24 hours. By contrast, the stations in New South Wales and Germany can change prices at any time; these data are aggregated into daily prices by taking either end-of-day values (New South Wales) or intraday averages (Germany). Intraday changes are relatively rare in New South Wales, and hence end-of-day values are representative of the actual transaction prices. In Germany, many stations change prices multiple times during the day, and 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.\textsuperscript{13} The Argus Media group’s OMR Oil Market Report collects daily regional wholesale prices and offers a database on a commercial basis.\textsuperscript{14}

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. The goal is to develop and compare the performance of multiple methods, and such contesting requires a common benchmark.

To establish a ground truth based on human recognition of cycles, eight RAs manually classified station-quarter observations.\textsuperscript{15} Each station $i$ in quarter $t$ is classified as either “cycling,” “maybe cycling,” or “not cycling.” The total number of manually labeled observations by region is 24,569 for Western Australia, 9,693 for New South Wales, and 35,685 for Germany. The RAs’ total working hours are approximately 260 for Western Australia, 210 for New South Wales, and 480 for Germany. The manual labeling of the data sets proceeded in three stages.

\textsuperscript{11} See FuelWatch (https://www.fuelwatch.wa.gov.au) and FuelCheck (https://www.fuelcheck.nsw.gov.au).

\textsuperscript{12} For details about the data, see Bundeskartellamt, Market Transparency Unit for Fuels (https://www.bundeskartellamt.de/EN/Tasks/markettransparencyunit_fuels/markettransparencyunit_fuels.html).

\textsuperscript{13} For average regional wholesale prices, see the Australian Institute of Petroleum (https://www.aip.com.au).

\textsuperscript{14} 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.

\textsuperscript{15} All research assistants are graduate or undergraduate students majoring in economics, mathematics, or statistics at Yale University.
Western Australia. First, all station-quarters in the Western Australia data were labeled by two RAs as a pilot project between July 2019 and June 2020. The first RA (a PhD student in economics) laid the groundwork by labeling approximately half of the Western Australia 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 Western Australia has one label based on the consensus of the two RAs.

New South Wales. The New South Wales data set is smaller than the Western Australia set 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 percent of the time. Subsequently, each RA independently labeled the entire data set in February–April 2021. Thus, each station-quarter \((i, t)\) in New South Wales carries three labels.

Germany. The same team of three RAs subsequently labeled a 5 percent random sample of the German data set in April–June 2021. In turn, these labels served as a source of the 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 percent random sample in June 2021. In total, 10 percent of the German data is triple labeled.

Risk of Collusion Is Low. In the computerized procedures for manually labeling data from New South Wales and Germany, each RA is given one randomly selected observation for labeling at a time. We believe that the risk of collusion among RAs is low because copying each other’s answers would require keeping records of random sequences of thousands of observations with their station-quarter identifiers, exchanging these long records, and 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 only once would be much easier.

Summary Statistics. Table 1 reports summary statistics. On the basis of these manual classification results, \(\text{Cycle}_{i,t}\) is defined as a binary variable indicating the presence of clear cycles. For Western Australia, each observation is labeled exactly once in accordance with the consensus of two RAs. The \(\text{Cycle}_{i,t}\) variable is equal to one if station-quarter \((i, t)\) is labeled cycling and zero otherwise. For New South Wales and Germany, which contain more ambiguous patterns, three RAs labeled each observation individually, and hence each station-quarter \((i, t)\) is triple labeled. The variable \(\text{Cycle}_{i,t}\) is equal to one for observations with triple cy-
### Table 1
Summary Statistics

|                  | Western Australia | New South Wales | Germany        |
|------------------|-------------------|----------------|----------------|
| **Sample period**| January 3, 2001–June 30, 2020 | August 1, 2016–July 31, 2020 | June 8, 2014–January 7, 2020 |
| Gasoline stations| 821               | 1,226          | 14,780         |
| Calendar quarters| 77                | 15             | 26             |
| Station-quarters | 25,463            | 9,693          | 353,086        |
| Labeled cycling by three research assistants | 0 (0.0) | 6,878 (71.0) | 14,116 (39.6) |
| Labeled cycling by two research assistants | 0 (0.0) | 906 (9.4)   | 7,173 (20.1)  |
| Labeled cycling by one research assistant | 15,007 (61.1) | 759 (7.8)  | 6,280 (17.6)  |
| Not labeled cycling by any research assistant | 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) |
| Total not manually labeled | 894 (100.0) | 0 (100.0)   | 317,401        |

**Note.** Each manually labeled station-quarter observation in the Western Australia data is classified by a single researcher as either “cycling,” “maybe cycling,” or “not cycling,” whereas the New South Wales and German data are triple labeled. Values in parentheses are percentages of the “total manually labeled” observations in each data set.
cling labels (that is, in accordance with three RAs’ unanimous decisions) and zero otherwise. Thus, the target for automatic detection is prepared in a relatively conservative manner.

3.3. Rationale for Daily Frequency and Quarterly Window

Choice of Sampling Frequency and Time Window. Several considerations led us to use the daily sampling frequency and the quarterly time window. First, we prioritized setting a common time frame for all three data sets. The goal is to compare the performance of various methods in multiple data sets under the same protocol; a detailed case study of any single region/country is not the main objective. The daily frequency is the finest granularity that can be commonly used across all data sets because retail prices in Western Australia are fixed for 24 hours by regulation (see Section 3.1). It is also the finest granularity used in most other studies (however, see the discussion of the German data).

Second, cyclicity implies repetition, the identification of which requires a sufficiently long time window. The existing studies on Western Australia and New South Wales report cycles with wavelengths of a week to several weeks, whereas those on Germany report both weekly and intraday cycles. The 12–13 weeks of a calendar quarter permit repeated observations of relatively long (for example, monthly) cycles.

Third, shorter-than-daily (for example, hourly) frequencies would be too costly for the research design, as systematic manual verification is its essential component. Visually inspecting and labeling a 10 percent subsample of the entire German data set 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 US$13.50, the total cost would be US$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 this 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.

Intraday Cycles in the German Data. We are aware of multiple studies that document intraday 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) during January 2007–June 2010

---

16 The sensitivity of the results under alternative criteria are assessed in Sections OB4–OB6 in Online Appendix OB.
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, intraday 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: one follows the other within 3 hours in 90 percent of the cases, and they are followed by three other major chains.

Given the well-documented presence of intraday cycles, one might wonder whether a focus on the daily data and multiday 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, these interesting short-run movements are lost. The choice of the daily frequency is driven by the comparative design of the research, which prioritizes the systematic comparisons across the three data sets and (costly) manual verification. Thus, researchers who wish to conduct an in-depth case study of the German fuel markets might want to analyze intraday patterns.

Nevertheless, the presence of shorter cycles does not preclude that of longer cycles; Bundeskartellamt (2011) confirms the existence of both. One should also note that the intraday cycles seem to follow a specific time schedule in which prices rapidly increase between 8:00 p.m. and midnight and gradually decrease from around 6:00 the following morning (Siekmann 2017). As Linder (2018) correctly points out, such a deterministic pattern is more consistent with intertemporal price discrimination than the Edgeworth cycles of Maskin and Tirole (1988; recall property 3—stochasticity—in Section 2.1). Hence, while interesting, the intraday cycles in Germany are outside the scope of this paper.

4. Models and Methods for Automatic Detection

This section explains how we formalize the four existing methods and the six new methods proposed. It also describes how the parameter values of each model are optimized.

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 versus Negative Runs. Castanias and Johnson (1993) compare the lengths of positive and negative changes. We formalize this idea by classifying each station-quarter \((i, t)\) as cycling \((\text{Cycle}_{i,t} = 1)\) if and only if

\[
\text{Mean}(\text{Len}(\text{Run}^+)) < \text{Mean}(\text{Len}(\text{Run}^-)) + \theta_{\text{PRNR}},
\]

where \(\text{Len}(\text{Run}^+)\) and \(\text{Len}(\text{Run}^-)\) denote the lengths of consecutive (multiday) price increases (and/or zero changes) and decreases within quarter \(t\), respectively.
The means are taken over these runs. The expression $\theta^{\text{PRNR}} \approx 0$ is a scalar threshold, which is treated as a parameter.\(^{17}\)

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

$$\left| \text{Mean}_{d \in t} (\Delta p_{i,d}^+) \right| > \left| \text{Mean}_{d \in t} (\Delta p_{i,d}^-) \right| + \theta^{\text{MIMD}},$$

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 $\theta^{\text{MIMD}} \approx 0$ is a scalar threshold. That is, a cycle is detected when the average price increase is greater than the average price decrease.\(^{18}\)

**Method 3: Negative Median Change.** Lewis (2009) identifies cycles if and only if

$$\text{Median}_{d \in t} (\Delta p_{i,d}) < \theta^{\text{NMC}},$$

where $\Delta p_{i,d}$ denotes a price change between days $d$ and $d - 1$, and $\theta^{\text{NMC}} \approx 0$ is a scalar threshold. In other words, the significantly negative median change is taken as evidence of price cycles.\(^{19}\)

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

$$\sum_{d \in t} \mathbb{I}\{\Delta p_{i,d} > \theta^{\text{MBPI}}_1\} \geq \theta^{\text{MBPI}}_2,$$

where $\mathbb{I}\{\}$ is an indicator function that equals one if the condition inside the brackets is satisfied and zero otherwise. The terms $\theta^{\text{MBPI}}_1$ and $\theta^{\text{MBPI}}_2$ are thresholds for big and many price increases, respectively. They set $\theta^{\text{MBPI}}_1 = 6$ (Australian cents per liter) and $\theta^{\text{MBPI}}_2 = 3.75$ (per quarter) in studying the Western Australia data.\(^{20}\) 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 this 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, and the third corresponds to a noncyclical regime.\(^{21}\)

---

\(^{17}\) 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.

\(^{18}\) 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.

\(^{19}\) Many subsequent studies use this method, including Bloch and Wills-Johnson (2010), Doyle, Muehlegger, and Samphantharak (2010), Lewis and Noel (2011), Lewis (2012), Eckert and Eckert (2014), Zimmerman, Yun, and Taylor (2013), and Byrne (2019). As a threshold for discretization, Lewis (2012) uses $-0.2$ US cents per gallon, whereas Doyle, Muehlegger, and Samphantharak (2010) and Zimmerman, Yun, and Taylor (2013) use $-0.5$ US cents per gallon.

\(^{20}\) Lewis (2009) also uses a similar method, with $\theta^{\text{MBPI}}_1 = 4$ (US cents/gallon) in a single day or 2 consecutive days.

\(^{21}\) Because these states are modeled as unobserved objects, using this approach as a definition is not straightforward. Zimmerman, Yun, and Taylor (2013) propose another definition that shares the spirit of Markov-switching regressions: Compare the probability that a price increase (decrease)
Detecting Edgeworth Cycles

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-the-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.

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 the procedures, Section OA1 in Online Appendix OA (and the replication package) provides additional details.

Method 5: Fourier Transform. 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 \( \Delta x \), the sample analog of the Fourier power spectrum (or periodogram) is

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

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 the 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 \( g_n \) (for more detail, see Section OA1 in Online Appendix OA).

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

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

where \( P_{i,t}(f) \) is the periodogram (equation [6]) of station-quarter \((i, t)\), and \( \theta_{\text{max}}^{\text{FT}} > 0 \) is a scalar threshold parameter.

is observed after two consecutive price increases (decreases), and if the conditional probability of a third consecutive increase is smaller than that of a third decrease, it is an indicator of cycles. We regard their approach as a variant of the Castanias and Johnson (1993) method. Finally, Noel (2019) defines the relenting and undercutting phases as consecutive days with cumulative increases and decreases of at least 3 Australian cents per liter, respectively, which is also close to the idea of Castanias and Johnson (1993).
Method 6: Lomb-Scargle Periodogram. The LS periodogram (Lomb 1976; Scargle 1982) characterizes periodicity in unevenly sampled time series.\(^{22}\) It has been extensively used in astrophysics because astronomical observations are subject to weather conditions and diurnal, lunar, or seasonal cycles. Formally, it is a generalized version of the classical periodogram (equation [6]):\(^{23}\)

\[
p_{\text{LS}}(f) = \frac{1}{2} \left[ \frac{\left( \sum_{n} n g_{n} \cos(2\pi f(x_{n} - \tau)) \right)^{2}}{\sum_{n} \cos^{2}(2\pi f(x_{n} - \tau))} + \frac{\left( \sum_{n} n g_{n} \sin(2\pi f(x_{n} - \tau)) \right)^{2}}{\sum_{n} \sin^{2}(2\pi f(x_{n} - \tau))} \right],
\]

where \(\tau\) is specified for each frequency \(f\) as

\[
\tau = \frac{1}{4\pi f} \tan^{-1} \left( \frac{\sum_{n} \sin(4\pi fx_{n})}{\sum_{n} \cos(4\pi fx_{n})} \right).
\]

We propose the following threshold condition to detect cycles:

\[
\max_{f} P_{\text{LS},it}(f) > \theta_{\text{max}}^{\text{LS}},
\]

where \(\theta_{\text{max}}^{\text{LS}} > 0\) is a scalar threshold parameter.

Method 7: Cubic Splines. 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, the discrete (daily) time series is smoothed by interpolating it with a commonly used continuous function.\(^{24}\) For each station-quarter \((i, t)\), CS is fit to its demeaned price series, \(\bar{p}_{i,t,d} = p_{i,t,d} - \text{Mean}_{d} (p_{i,d})\), and the number of times the fitted function \(\text{CS}_{i,t}(d)\) crosses the \(d\)-axis (that is, equals zero) is counted. Operationally, the number of real roots is counted, and cycles are detected with the condition

\[
\text{Roots} (\text{CS}_{i,t}(d)) > \theta_{\text{Root}}^{\text{CS}},
\]

where \(\theta_{\text{Root}}^{\text{CS}} > 0\) 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. Recurrent neural networks with LSTM (Hochreiter and Schmidhuber 1997) are a class of artificial neural network (ANN) models for sequential data. These LSTM networks have become a de facto standard for recognizing and predicting complicated patterns in many applica-

\(^{22}\) The data are evenly sampled at a daily frequency and can be analyzed by the Fourier transform (FT) method alone, but the Lomb-Scargle periodogram offers additional benefits. One is conceptual: it is interpretable as a kind of nonparametric regression (see method 6 in Section OA1 in Online Appendix OA). Another is practical: its off-the-shelf computational implementation can offer more granular periodograms.

\(^{23}\) Section OA1 in Online Appendix OA (method 6) explains how this expression relates to FT.

\(^{24}\) We use a cubic Hermite interpolator, which is a spline where each piece is a third-degree polynomial of Hermite form. Section OA1 in Online Appendix OA (method 7) explains the details of this functional form.
tions, including speech, handwriting, language, and polyphonic music. Because LSTM is relatively new, we explain this method in greater detail.

Econometrically speaking, LSTM is a nonparametric model for time-series analysis. It is a recursive dynamic model whose behavior centers on a collection of pairs of \( B_l \times 1 \) vector-valued latent state variables, \( s^l_d \) and \( c^l_d \), where \( l = 1, 2, \ldots, L \) is an index of layers. As this notation suggests, we use a multilayer architecture (also known as deep neural networks) to enhance the model’s flexibility. The term \( B_l \) represents the number of blocks per layer, which are analogous to neurons (basic computing units) in other ANN models. The term \( s^l_1 \) is an output state that represents the current, short-term state, whereas \( c^l_d \) is called a cell state and retains long-term memory. The latter is designed to capture lagged dependence between the state and input variables, and thereby plays the role of a memory cell in electronic computers.

These state variables evolve according to the following Markov process:

\[
s^l_d = \tanh(c^l_d) \circ \Lambda(\omega^l_1 + \omega^l_2 \Delta p_d + \omega^l_3 s^{l-1}_d), \tag{12}
\]
\[
c^l_d = \tanh(\omega^l_4 + \omega^l_5 \Delta p_d + \omega^l_6 s^{l-1}_d) \circ \Lambda(\omega^l_7 + \omega^l_8 \Delta p_d + \omega^l_9 s^{l-1}_d)
+ c^{l-1}_d \circ [1 - \Lambda(\omega^l_4 + \omega^l_5 \Delta p_d + \omega^l_6 s^{l-1}_d)], \tag{13}
\]

where \( d = 1, 2, \ldots, D \) is the index of days; \( \Delta p_d \equiv p_d - p_{d-1} \) (where \( \Delta p_1 = 0 \)); \( \tanh(x) \equiv (e^x - e^{-x})/(e^x + e^{-x}) \) is the hyperbolic tangent function; the symbol \( \circ \) denotes the Hadamard (element-wise) product; and \( \Lambda(x) \equiv e^x/(1 + e^x) \) is the cumulative distribution function of the logistic distribution. The \( \omega \) terms are weight parameters with the following dimensionality: \( \omega^l_1, \omega^l_2, \omega^l_3, \omega^l_4, \omega^l_5, \omega^l_6, \) and \( \omega^l_9 \) are \( B_l \times 1 \) vectors, and \( \omega^l_7, \omega^l_8, \) and \( \omega^l_9 \) are \( B_l \times B_{l-1} \) matrices. Thus, \( B \equiv (B_1, B_2, \ldots, B_L) \) 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, the terms \( s^{l-1}_d, c^{l-1}_d \), and \( B_{l-1} \) in equations (12) and (13) should be replaced by \( s^L_{d-1}, c^L_{d-1} \), and \( B_L \) when \( l = 1 \). After the final layer \( L \) of the last day \( D = 90 \) of quarter \( t \), cycles in station-quarter \((i, t)\) are detected if and only if

\[
s^A(p_{i,t}; \theta^{LSTM}) \equiv \omega_{10} + \omega_{11} s^L_D > 0, \tag{14}
\]

where \( \omega_{10} \) is a scalar; \( \omega_{11} \) is a \( B_L \times 1 \) vector; and \( \theta^{LSTM} \equiv (\omega, L, B) \) collectively denotes all parameters, including the many weights in \( \omega \equiv (\omega^l_1, \omega^l_2, \ldots, \omega^l_9)_{l=1}^L, \omega_{10}, \omega_{11} \), the number of layers \( L \), and the profile of the number of blocks in each layer \( B \).

We set \( L = 3 \) and \( B = (16, 8, 4) \) and find the value of \( \omega \) that approximately maxi-

---

25 Except for the multilayer design, this specification mostly follows Greff et al. (2017), in which one of the original proponents of long short-term memory (LSTM) and his team compare many of its variants and show that their simple vanilla specification outperforms others.

26 See Section OA1 in Online Appendix OA (method 8) for further details on this specification and computational implementation.
mizes the accuracy of the prediction (explained in Section 4.3 and Section OA2 in Online Appendix OA).

In summary, LSTM sequentially processes the daily price data in a flexible Markov model with many latent states and parameters. The terminal state $s^*$ plays the role of a latent score to detect cycles.

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

$$g^m_{i,t} \equiv \text{LHS}^m_{i,t} - \text{RHS}^m_{i,t} \quad (15)$$

denote a gap, the scalar difference between the left-hand side and the right-hand side of the inequality that defines each method $m = 1, 2, \ldots, M$, excluding the threshold parameter $\theta^m$. For example, inequality (3) defines method 2. Hence, $g^2_{i,t} = |\text{Mean}_d(\Delta p_{i,d}^+) - |\text{Mean}_d(\Delta p_{i,d}^-)|.27$ Let

$$g_{i,t} \equiv (g^m_{i,t})_{m=1}^M \quad (16)$$

denote their vector, where $M = 7.28$ We construct a decision-tree classification algorithm that takes values from $g_{i,t}$ as inputs and predicts $\text{Cycle}_{i,t} = 1$ if and only if

$$h(g_{i,t}; \omega^{RF}, \kappa^{RF}) \equiv \sum_{k=1}^K \omega^k_{g} \, 1\{g_{i,t} \in R_k\} \equiv \sum_{k=1}^K \omega^k_{g} \, 1\{h(g_{i,t}; \kappa^{RF}) > 0, \quad (17)$$

where $K$ is the number of adaptive basis functions, $\omega^k_{g}$ is the weight of the $k$th basis function, $R_k$ is the $k$th region in the $M$-dimensional space of $g_{i,t}$, and $\kappa^{RF}$ encodes both the choice of variables (elements of $g_{i,t}$) and their threshold values that determine region $R_k$.29 Because finding the truly optimal partitioning is a computationally difficult (combinatorial) problem, we use an RF algorithm to stochastically approximate it (for details, see Section OA1 in Online Appendix OA [method 9]). Thus, this method aggregates and generalizes methods 1–7 in a flexible manner that permits multiple thresholds and interactions between $g^m_{i,t}$ terms. We denote its full set of parameters as $\theta^{RF} \equiv (\omega^{RF}, \kappa^{RF}) \equiv ((\omega^k_{g})_{k=1}^K, (\kappa^k_{g})_{k=1}^K)$.

Method 10: Ensemble in Long Short-Term Memory. This method combines methods 1–8 within an extended LSTM by incorporating $g_{i,t}$ in expression (16) as an additional variable in the laws of motion:

$$s^l_d = \tanh(c^l_d) \circ \Lambda(\omega^l_1 + \omega^l_2 \Delta p_d + \omega^l_3 s^{l-1}_d + \omega^l_4 g) \quad (18)$$

and

$$c^l_d = \tanh(\omega^l_1 + \omega^l_2 \Delta p_d + \omega^l_3 s^{l-1}_d + \omega^l_4 g) \circ \Lambda(\omega^l_1 + \omega^l_3 \Delta p_d + \omega^l_4 s^{l-1}_d + \omega^l_5 g) + c^{l-1}_d \circ [1 - \Lambda(\omega^l_1 + \omega^l_2 \Delta p_d + \omega^l_3 s^{l-1}_d + \omega^l_4 g)], \quad (19)$$

27 Methods 1–3 and 5–7 are one-parameter models like this example. For method 4, we define $g^4_{i,t} = \sum_{d,t} I\{\Delta p_{i,d} > \theta^{MBPI}_{i,t}^*\}$, where $\theta^{MBPI}_{i,t}^*$ is the accuracy-maximizing value of $\theta^{MBPI}_{i,t}$.

28 The computational implementation incorporates two additional variants of (each of) methods 5–7, which are explained in Section OA1 in Online Appendix OA (method 9). Hence, the eventual value of $M$ is $7 + (2 \times 3) = 13$.

29 See Murphy (2012, chap. 16) for an introduction to adaptive basis-function models including random forests.
where \( \omega_{i2}^l, \omega_{i3}^l, \) and \( \omega_{i4}^l \) are \( B_i \times M \) matrices of weight parameters for \( g_{i,t} \) (the station-quarter subscripts are suppressed in equations [18] and [19]). Other implementation details are the same as in method 8.

### 4.3. Optimization of Parameter Values (Training)

**Accuracy Maximization.** Whereas the existing research typically calibrates (that is, manually selects) the threshold parameters, we optimize this process by choosing the parameter values that maximize accuracy, which are defined as the percentage of correct predictions,

\[
\text{Percentage Correct} (\theta) = \frac{\sum_{(i,t)} \mathbb{1}\{\text{Cycle}_{i,t}(\theta) = \text{Cycle}_{i,t}\}}{\text{All Predictions}} \times 100,
\]

where \( \text{Cycle}_{i,t}(\theta) \in \{0, 1\} \) is the algorithmic prediction for observation \((i, t)\) at parameter value \( \theta \), and \( \text{Cycle}_{i,t} \in \{0, 1\} \) is the manual classification label (data). Two types of prediction errors, false negative and false positive, are analogously defined in Section OA2 in Online Appendix OA. Thus,

\[
\theta^* \equiv \arg \max_\theta \text{Percentage Correct} (\theta)
\]

characterizes the optimized (or trained) model for each method.

**Splitting Data into Training and Testing Subsamples.** We optimize and evaluate each method. This process is done separately for each of the three data sets (Western Australia, New South Wales, and Germany):

1. Randomly split each labeled data set into an 80 percent training subsample and a 20 percent testing subsample.
2. Optimize the parameter values of each model in the 80 percent training subsample.
3. Assess the out-of-sample prediction accuracy of the trained model in the 20 percent testing subsample.\(^{30}\)
4. Repeat these three steps 101 times.\(^{31}\)
5. Report the medians of the optimized parameter values and the medians and standard deviations of the prediction-accuracy results.

### 5. Results

Table 2 summarizes the performance of all methods for each data set. Reported are the median accuracy, the composition of correct and incorrect predictions, and the associated parameter value(s) \( \theta^* \) for each method.

**Western Australia.** For Western Australia, clear-cut cycles of deterministic frequencies are known to exist. Table 2 shows that almost all methods achieve

\(^{30}\)This cross-validation procedure is particularly important for the nonparametric models of methods 8–10, which contain many parameters and could potentially overfit the training subsample.

\(^{31}\)An odd number of bootstrap sample splits facilitates the selection of the medians in step 5.
|                  | PRNR     | MIMD     | NMC      | MBPI     | FT     | LS      | CS      | LSTM    | E-RF    | E-LSTM  |
|------------------|----------|----------|----------|----------|--------|---------|---------|---------|---------|---------|
|                  | (1)      | (2)      | (3)      | (4)      | (5)    | (6)     | (7)     | (8)     | (9)     | (10)    |
| Western Australia ($N = 24,569$): |          |          |          |          |        |         |         |         |         |         |
| Parameter 1      | −1.16    | 6.13     | −.20     | 5.05     | .12    | .21     | 22.50   |         |         |         |
| Parameter 2      |          |          |          |          | 5      |         |         |         |         |         |
| Accuracy rank    |          |          |          |          | 5      | 4       | 9       | 6       | 8       | 7       |
| Percentage Correct | 90.80    | 91.27    | 89.34    | 90.23    | 90.11  | 90.15   | 85.47   | 99.25   | 99.04   | 99.25   |
|                  | (.37)    | (.38)    | (.38)    | (.36)    | (.40)  | (.36)   | (.45)   | (.18)   | (.15)   | (.14)   |
| Percentage correctly labeled cycling | 55.27    | 55.70    | 57.08    | 60.74    | 58.24  | 57.92   | 56.41   | 60.62   | 60.97   | 60.34   |
| Percentage correctly labeled not cycling | 35.53    | 35.57    | 32.25    | 29.49    | 31.87  | 32.23   | 29.06   | 38.62   | 38.07   | 38.91   |
| Percentage false negative | 5.27     | 5.27     | 3.34     | .71      | 2.48   | 3.30    | 5.29    | .35     | .61     | .31     |
| Percentage false positive | 3.93     | 3.46     | 7.33     | 9.06     | 7.41   | 6.55    | 9.24    | .41     | .35     | .45     |
| New South Wales ($N = 9,693$): |          |          |          |          |        |         |         |         |         |         |
| Parameter 1      | 4.20     | 5.76     | 1.01     | 14.90    | .20    | .57     | 4.50    |         |         |         |
| Parameter 2      |          |          |          |          | 2      |         |         |         |         |         |
| Accuracy rank    |          |          |          |          | 7      | 8       | 10      | 4       | 6       | 5       |
| Percentage Correct | 78.55    | 78.39    | 70.96    | 81.59    | 80.71  | 80.82   | 73.90   | 89.63   | 87.42   | 90.30   |
|                  | (.85)    | (.88)    | (.97)    | (.86)    | (.80)  | (.80)   | (.89)   | (.67)   | (.69)   | (.67)   |
| Percentage correctly labeled cycling | 67.04    | 65.09    | 70.96    | 64.62    | 66.53  | 66.43   | 70.40   | 67.20   | 67.10   | 65.60   |
| Percentage correctly labeled not cycling | 11.50    | 13.31    | .00      | 16.97    | 14.18  | 14.39   | 3.51    | 22.43   | 20.32   | 24.70   |
| Percentage false negative | 3.30     | 4.85     | .00      | 6.55     | 5.47   | 4.02    | .77     | 4.33    | 8.35    | 2.99    |
| Percentage false positive | 18.15    | 16.76    | 29.04    | 11.86    | 13.82  | 15.16   | 25.32   | 6.03    | 4.23    | 6.70    |
| Germany ($N = 35,685$): |          |          |          |          |        |         |         |         |         |         |
| Parameter 1      | −3.48    | .30      | −.45     | 1.25     | .24    | .62     | 24.50   |         |         |         |
| Parameter 2      |          |          |          |          | 14     |         |         |         |         |         |
| Accuracy rank    |          |          |          |          | 9      | 6       | 7       | 5       | 8       | 10      |
| Percentage Correct | 60.38    | 60.61    | 60.53    | 65.39    | 60.50  | 60.36   | 71.28   | 74.61   | 76.14   | 79.58   |
|                  | (.49)    | (.50)    | (.52)    | (.52)    | (.56)  | (.59)   | (.42)   | (.44)   | (.14)   | (.53)   |
| Percentage correctly labeled cycling | .00      | 1.25     | .07      | 14.77    | .00    | .00     | 25.88   | 23.46   | 23.96   | 29.96   |
| Percentage correctly labeled not cycling | 60.38    | 59.37    | 60.46    | 50.62    | 60.50  | 60.36   | 45.40   | 51.16   | 52.18   | 49.63   |
| Percentage false negative | 39.62    | 38.07    | 39.40    | 24.65    | 39.50  | 39.57   | 14.28   | 15.99   | 15.75   | 9.50    |
| Percentage false positive | .00      | 1.32     | .07      | 9.96     | .00    | .07     | 14.45   | 9.40    | 8.11    | 10.91   |

Note. Columns 8–10 do not report parameter values because there are too many to list. The sample was randomly split into an 80 percent training subsample and a 20 percent testing subsample 101 times. In each split, the former subsample is used for setting parameter values; their medians are reported. The accuracy statistics are medians from the 101 testing subsamples as well; their standard deviations are in parentheses. CS = cubic splines; E-LSTM = ensemble in long short-term memory; E-RF = ensemble in random forests; FT = Fourier transform; LS = Lomb-Scargle periodogram; LSTM = long short-term memory; MBPI = many big price increases; MIMD = mean increase versus mean decrease; NMC = negative median change; PRNR = positive runs versus negative runs.
high accuracy near or above 90 percent. The flexible, nonparametric models of methods 8–10 do particularly well at above 99 percent accuracy.

Some parameter values are informative about the underlying data patterns. For example, CS lags behind all other methods with (a still respectable) 85 percent accuracy. Its parameter value, $\theta_{\text{Roots}}^{\text{CS}} = 22.5$, suggests that the model is trained to focus on shorter cycles with wavelengths less than $90/(22.5/2) = 8$ days. Byrne and de Roos (2019) show that both 1- and 2-week cycles exist in Western Australia. Thus, the inferior performance of CS stems from missing the latter, longer cycles.

Another interesting result concerns MBPI, which achieves 90 percent accuracy. Byrne and de Roos (2019) set $\theta_1^{\text{MBPI}} = 6$ and $\theta_2^{\text{MBPI}} = 3.75$ in their original study of Western Australia. 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.

**New South Wales.** Cycle detection in New South Wales is not as easy as in Western Australia, but most methods achieve near or above 80 percent accuracy. The nonparametric methods are top performers again (87–90 percent), followed by MBPI and the spectral methods (81–82 percent). By contrast, CS (74 percent) and NMC (71 percent) 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 Western Australia. 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 percent). This finding alone does not necessarily invalidate the use of NMC in other data sets but cautions against overly relying on any single metric.

**Germany.** Most methods fail in Germany, where cycles are subtler and data are noisier (that is, the RAs reach unanimous decisions less often). The only method that achieves accuracy near 80 percent is E-LSTM, and it is followed in accuracy by E-RF (76 percent) and LSTM (75 percent). Somewhat surprisingly, CS (71 percent) outperforms all other parametric models; MBPI (65 percent) is the only method from the existing literature with nondegenerate predictions, presumably because it does not exclusively rely on asymmetry.

---

32 As Table 1 shows, 71 percent of the New South Wales data is unanimously labeled cycling by three research assistants (RAs), whereas $9.4 + 7.8 = 17.2$ percent is labeled as such by only one RA or two RAs. In the German sample, only 39.6 percent is unanimously labeled cycling, whereas RAs disagree about $20.1 + 17.6 = 37.7$ percent of the data. Section OB4 in Online Appendix OB reports results based on subsamples that eliminate such observations with disagreements. By contrast, Section OB5 in Online Appendix OB investigates how the algorithms classify ambiguous observations (that is, station-quarters on which RAs disagree and/or choose “maybe cycling”). Section OB6 in Online Appendix OB examines such heterogeneity by labelers in detail.
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 cyclicity in fuzzier manners (methods 4 and 7) manage to make at least some correct (nondegenerate) predictions. These results suggest that not all of the German cycles conform to the idealized patterns of asymmetry or cyclicity and that less rigid classification rules could be relatively more robust to irregular patterns and noise.

The parameter values of $CS_{\text{Roots}} = 24.50$ and $MBPI_{\text{q}} = 14$ suggest that the German cycles are approximately weekly. That is, $CS_{\text{Roots}} = 24.50$ means at least as many ups as downs are often recorded in cycling observations, which translate into the wavelength of $90/(24.5/2) = 7.3$ days or shorter. Similarly, $MBPI_{\text{q}} = 14$ requires at least as many large jumps within a calendar quarter and hence implies the wavelength of $90/14 = 6.4$ 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 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 with 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: it is 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. 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 (that is, permitting their interactions and multiple thresholds).

**Performance on Simulated Cycles.** Section OA3 in Online Appendix OA includes an examination of the performance of the 10 methods 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 data-generating process (DGP) and deploy the three pretrained versions (Western Australia, New South Wales, 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 that the real-world data are qualitatively different from simulated data with artificial cycles.
6. How Much Data Is Needed?

The results of the accuracy contest in Section 5 show that more flexible methods tend to outperform simple parametric ones, which is not surprising. The real question is whether the cost of training complicated machine-learning algorithms, which are known to require a lot of data, is acceptable. 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 data set. The rate of improvement is different across methods however. Figure 2 shows performance when the training data set is restricted. The exact numbers underlying these plots are reported in Table OB3 in Online Appendix OB.

Methods 1–7 and 9 perform surprisingly well with only .1 percent of the data, which corresponds to 25, 10, and 36 observations in Western Australia, New South Wales, 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.

It is not surprising that simple models with a parameter or two (methods 1–7) require only a few dozen observations. All that is needed to distinguish cycles from noncycles is to adjust a numerical threshold or two. 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 underutilize.

Methods 8 and 10 contain a few thousand parameters and obviously need more data. For instance, E-LSTM's accuracy in New South Wales is below 50 percent when it uses only 10 observations (.1 percent subsamples). Fortunately, their performance dramatically improves with a mere 1 percent subsample, and they start outperforming all other methods when 5 percent subsamples are used. The critical sample size above which they perform the best is on the order of several hundred observations. The associated cost of manual labeling is only tens of RA hours, or a few hundred US dollars. Thus, even though LSTM and E-LSTM require more data for a given accuracy level, their total cost is surprisingly low, which makes them the highest accuracy methods even for 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

---

33 Strictly speaking, ensemble in random forests slightly outperforms ensemble in LSTM in subsamples up to 40 percent in Western Australia, although their mean differences are small relative to their standard deviations (see Tables OB3 and OB4 in Online Appendix OB).

34 Table OB3 in Online Appendix OB reports the total cost of manual labeling for each data set. 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.
Figure 2. Gains from additional data. A, Western Australia; B, New South Wales; C, Germany.
that humans recognize are relatively simple after all, even though explicitly articulating them might be difficult.

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 (Deltas 2008; Clark and Houde 2014; Byrne 2019), whereas others find the opposite relationships (Lewis 2009; Zimmerman, Yun, and Taylor 2013; Noel 2015). 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 noncycling observations. The first column is based on the manual classification and serves as a benchmark. The mean margins in cycling and noncycling observations in Western Australia are AU¢11.86 and AU¢9.47, respectively. The mean difference is AU¢2.39. The t-test (based on Welch’s t-statistic) rejects the null hypothesis that the difference in means is zero at the .1 percent significance level. Hence, price cycles are positively correlated with margins in Western Australia. The same analysis yields similar results in New South Wales.

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 the region.

Algorithmic Cycle Detection and Margins. In Western Australia, all algorithmic methods reach the same conclusion that margins are higher in cycling observations. Broadly similar results also emerge in New South Wales, even though one method fails (method 3) and one reaches the opposite conclusion (method 7). These discrepancies suggest that researchers will find a positive or negative cycle-margin relationship depending on the operational definition of cycles.

The analysis of the German data highlights this point even more vividly. Both the manual classification and methods 7–10 suggest significantly negative rela-

\footnote{The measure of profit margin is the difference between the retail price and the wholesale price before tax, as defined in equation (1), in Australian cents in Western Australia and New South Wales and euro cents in Germany. Note that 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.}

\footnote{Determining the exact source of heterogeneity is beyond the scope of this paper. There can be many reasons, and Edgeworth cycles are only one possible mechanism. Our purpose is to illustrate with concrete examples how different methods could lead to different findings and policy implications.}
| Table 3 | Profit Margins by Cycle Status |
|---------|--------------------------------|
|         | Manual | PRNR | MIMD | NMC | MBPI | FT  | LS  | CS  | LSTM | E-RF | E-LSTM |
| Western Australia (N = 24,569):  |        |      |      |     |      |     |     |     |      |      |        |
| Cycling:                           |        |      |      |     |      |     |     |     |      |      |        |
| Observations                       | 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 |
| SD                                  | 4.01   | 3.80  | 3.74  | 3.98  | 4.13  | 3.87  | 3.85  | 4.04  | 4.01  | 4.01  | 4.01  |
| Not cycling:                        |        |      |      |     |      |     |     |     |      |      |        |
| Observations                       | 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  |
| SD                                  | 4.97   | 5.04  | 4.98  | 5.22  | 5.20  | 5.18  | 5.03  | 5.02  | 4.97  | 4.97  | 4.96  |
| Difference                          |        |      |      |     |      |     |     |     |      |      |        |
| Mean difference                     | 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-statistic                | 39.53  | 46.74 | 53.80 | 32.96 | 25.64 | 43.53 | 50.02 | 38.67 | 39.53 | 39.55 | 39.60 |
| Degrees of freedom                 | 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 |
| New South Wales (N = 9,693):       |        |      |      |     |      |     |     |     |      |      |        |
| Cycling:                           |        |      |      |     |      |     |     |     |      |      |        |
| Observations                       | 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 |
| SD                                  | 5.51   | 5.80  | 5.58  | 6.04  | 5.48  | 5.89  | 5.84  | 5.99  | 5.53  | 5.53  | 5.56  |
| Not cycling:                       |        |      |      |     |      |     |     |     |      |      |        |
| Observations                       | 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 | 10.25 | 10.64 | 10.33 |
| SD                                  | 7.10   | 7.31  | 7.01  | 6.92  | 6.56  | 6.85  | 6.79  | 7.01  | 7.08  | 7.08  | 7.08  |
|                  | Mean difference | Mean difference | Welch’s $t$-statistic | Degrees of freedom |
|------------------|-----------------|-----------------|-----------------------|--------------------|
|                  | 1.27            | .48             | 8.50                  | 4,266              |
|                  | 4.02            | 3.30            | 21.94                 | 1,663              |
| Welch’s $t$-statistic | 21.24          | 2.97            | 4.70                  | 2,106              |
|                  | 4.02            | 3.30            | 21.24                 | 1,663              |
| Degrees of freedom | 4,266           | 1,663           | 2,106                 | 3,423              |
|                  | 72              | 114.11          | 32.10                 | 2,252              |
|                  | 8,763           | 115.64          | 31.40                 | 472                |
|                  | 7               | 98.19           | 3.60                  | 2,870              |
|                  | 7               | 98.38           | 3.60                  | 2,870              |
|                  | 14,281          | 98.19           | 3.60                  | 2,870              |
|                  | 11,762          | 98.38           | 3.19                  | 2,870              |
|                  | 13,574          | 98.16           | 3.59                  | 2,870              |
|                  | 15,299          | 98.18           | 3.51                  | 2,870              |
| $p$-Value        | <.001           | <.001           | <.001                 | <.001              |
| Germany $N = 35,685$ |                |                 |                       |                    |
| Cycling:         |                 |                 |                       |                    |
| Observations     | 14,116          | 0               | 1,013                 | 14,281             |
|                  | 10,13           | 114.11          | 32.10                 | 11,762             |
|                  | 7               | 98.19           | 3.60                  | 32.10              |
|                  | 8,763           | 115.64          | 31.40                 | 31.40              |
|                  | 7               | 98.38           | 3.60                  | 31.40              |
|                  | 7               | 98.19           | 3.60                  | 31.40              |
|                  | 14,281          | 98.19           | 3.60                  | 31.40              |
|                  | 11,762          | 98.38           | 3.59                  | 31.40              |
|                  | 13,574          | 98.16           | 3.51                  | 31.40              |
|                  | 15,299          | 98.18           | 3.51                  | 31.40              |
| $p$-Value        | <.001           | <.001           | <.001                 | <.001              |
| Not cycling:     |                 |                 |                       |                    |
| Observations     | 21,569          | 35,685          | 34,672                | 21,404             |
|                  | 34,672          | 35,678          | 26,922                | 23,923             |
|                  | 35,613          | 35,678          | 35,678                | 21,111             |
|                  | 26,922          | 35,678          | 21,404                | 20,386             |
|                  | 35,685          | 35,678          | 23,923                | 20,386             |
|                  | 35,613          | 35,678          | 21,404                | 20,386             |
|                  | 26,922          | 35,678          | 23,923                | 20,386             |
|                  | 35,685          | 35,678          | 21,404                | 20,386             |
|                  | 35,613          | 35,678          | 23,923                | 20,386             |
| $p$-Value        | <.001           | <.001           | <.001                 | <.001              |

**Note.** The median-accuracy version of each method is used. The unit of measurement (of means and standard deviations) is the Australian cent in Western Australia and New South Wales and the euro cent in Germany. 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. CS = cubic splines; E-LSTM = ensemble in long short-term memory; E-RF = ensemble in random forests; FT = Fourier transform; LS = Lomb-Scargle periodogram; LSTM = long short-term memory; MBPI = many big price increases; MIMD = mean increase versus mean decrease; NMC = negative median change; PRNR = positive runs versus negative runs.
relationships between cycles and margins, but methods 2–6 lead to positive mean differences. These positive differences are highly statistically significant in methods 2–4. Some of them entail degenerate predictions (see Section 5), but method 4 features reasonable parameter values and achieves at least 65 percent accuracy. Hence, we cannot dismiss these discrepancies as purely random anomalies.

Margins and Asymmetric Cycles. Note that the classification so far has focused on cyclicity but not asymmetry. One might wonder whether the findings could change if asymmetric cycles are studied specifically. The answer is no. The results are virtually the same with asymmetric cycles.

Table 4 compares the mean differences of margins between the manual benchmark (from Table 3) and its refined version in which asymmetry (based on the negative median change, \( \text{Median}_{\Delta p_{i,c}} < 0 \)) is required 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. Even though operational definitions would appear to be a matter of technical details, they could even dictate the policy implications of empirical research on Edgeworth cycles.

7.2. Additional Findings

The results in Sections 5, 6, and 7.1 constitute the main findings, but the curious patterns in Section 7.1 present additional puzzles. We address them in the following discussion and report supporting evidence in Online Appendix OC.
Why Existing Methods Work in Australia but Fail in Germany. Most of the cycles in Australia follow specific (almost deterministic) wavelengths and exhibit strong asymmetry, whereas the cycles in Germany are noisier and not always asymmetric (see supplementary figures in Online Appendix OC). The existence of asymmetric noncycles in Germany further complicates the issue. Hence, asymmetry-based methods correctly identify cycles in Australia but not in Germany.

Why Margins and Cycles Correlate Positively in Australia but Negatively in Germany. In all data sets, 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). Volatility and cyclicality are correlated positively in Australia but negatively in Germany (see Online Appendix OC for supplementary figures). Therefore, the average level and cyclicality of margins are correlated positively in Australia but negatively in Germany.

How Can Cycles Be Less Volatile than Noncycles? Cyclicality implies systematic—but not necessarily large—movements; not all large/frequent movements follow cycles. Many German observations exhibit high volatility without any discernible pattern, which explains the existence of volatile noncycles in the data.

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 Online Appendix OC for supplementary figures). In Germany however volatility is a poor predictor of cyclicality.

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

Nevertheless, if the intraday cycles follow the wavelength of exactly 24 hours (or any number that can divide 24 evenly), they would be averaged out in the process of computing daily prices and would not affect the observations. The existing studies suggest that they do follow exactly 24-hour cycles (see Section 3.3). Hence, how intraday cycles affect the multidaily volatility in our data is not obvious.

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: it is the second-best option. If a perfect mathematical definition existed, no detection problem would arise in the first place. In the absence of such a formula, the research relies on rules of thumb val-

---

37 One possibility is the existence of medium-wavelength 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.
idated by selective eyeballing. 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 data set. 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? Figure 3 plots the fractions of stations that exhibit price cycles in Australia and Germany. Throughout this section, the recognition of cycles is based on the median-performance version (parameter values) of the most accurate algorithm (method 10), which is applied to the entire data set—both labeled and unlabeled—in each region.

Western Australia and New South Wales show mostly high percentages of price cycling. Western Australia offers the longest data period. Byrne and de Roos (2019) document clear price cycles in two subperiods (2007–8 and 2010–15), both of which correspond to the periods in which cycles are prevalent according to our method. Thus, the results of our method confirm Byrne and de Roos’s description of the Western Australia data in terms of time series. The New South Wales data set starts relatively recently in quarter 4 of 2016. Its range of approximately 70–90 percent is comparable to that of Western Australia.

Germany shows greater heterogeneity across its 10 geographic zones (Postleitzonen [LZ]). Former East Germany is represented by LZ0 and LZ1, LZ2–LZ6 are northwestern regions, and LZ7–LZ9 roughly correspond to the southern states of Baden-Württemberg and Bavaria. Three patterns emerge. First, whereas results for 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 percent in 2015–17 to 0–70 percent in 2018–19. The timing of this change seems to roughly coincide with the introduction and dissemination of automatic pricing algorithms (see Assad et al. 2024), but the clarification of their causal relationship would require further research. Section OD1 in Online Appendix OD includes a similar investigation of the relationship between macroeconomic shocks and price cycles, which seems complex.

Readers might wonder what causes sudden increases and decreases in Western Australia in the 2000s. Some of them reflect genuine changes in the number of stations with price cycles; others could be due to noise in the original data because the database lacks a consistent station identifier. Even though we tried to reconstruct the most balanced panel data possible (using street addresses and other observable characteristics), the recorded number of stations varies across time, sometimes quite dramatically.

For maps and further details, see Wikipedia, Postal Codes in Germany (https://en.wikipedia.org/wiki/Postal_codes_in_Germany).
Figure 3. Percentage of stations with price cycles
Spatial Patterns. The geographical scope of price cycles (for example, local, citywide, or regional) and their synchronization patterns might shed light on their mechanism. They could potentially inform the definition of relevant markets for antitrust purposes as well. 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.

The measure of correlation between stations is constructed as follows. First, within each region all possible pairs of stations are split into seven distance bins (< 1, 1–5, . . . , 50–100, >100 kilometers) according to their Euclidean distances. Second, for each pair the percentage of quarters in which their cycle statuses match (that is, either both stations exhibit cycles or neither of them does) is calculated. Third, the average of these percentages, either across all pairs or across pairs of same-brand stations, is calculated for each distance bin in each region. 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.

Figure 4 reports the spatial patterns of correlation in four graphs, and four patterns emerge. First, the majority of the station-pair-quarter observations share a cycle status, with the exception of the most distant (>100 kilometers) bin in rural New South Wales. Second, the cities and the rural areas of Australia exhibit qualitatively different patterns. The station pairs within Perth and Sydney (the capital cities of Western Australia and New South Wales, respectively) tend to show high correlations with limited variability across distance bins, whereas the rest of New South Wales features correlations that decrease with distance. 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 pairs in both Australia and Germany, especially in the 0–10 kilometer bins in Germany.

Section OD2 in Online Appendix OD includes additional investigations of how the relationship between cycle correlation and distance varies by time period in Western Australia. No such analysis is possible in New South Wales and Germany because of relatively short sample periods in these data sets.

8. Practical Recommendations

On the basis of the findings in Sections 5–7, we suggest the following 10 steps as a practical guide for automating the detection of Edgeworth cycles. They are not necessarily the most rigorous or precise procedures in any technical way. The

\[40\] Note that we consider only pairs that have at least 12 calendar quarters of valid data in common.

\[41\] We put “correlation” in quotes because we use the percentage of quarters with matched cycle statuses instead of the correlation coefficient, which is undefined when a station always (or never) shows cycles.

\[42\] All of the Western Australia stations (with sufficient observations for these plots) are in Perth, which is why we do not split the data for Western Australia into urban and rural areas as with the data for New South Wales.
Detecting Edgeworth Cycles

The analyst should use their own judgment to address unique challenges posed by each new data set.

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

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

3. As a first attempt to algorithmically distinguish cycles from noncycles, 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.

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 than the one assumed here) might have to be treated with the same care as in the selection and training of jury in trials. Online Appendix OE discusses this issue in detail.

Figure 4. Correlation of cycles between stations. A, Australia: all pairs; B, Australia: same-brand pairs only; C, Germany: all pairs; D, Germany: same-brand pairs only.
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), which are listed 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 statistic for median price change from method 3 (NMC) offers a simple way to capture asymmetry. For example, one can distinguish between Edgeworth-type asymmetry (that is, the median change is negative), inverse-Edgeworth asymmetry (that is, the median change is positive), and symmetry (that is, 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 (for example, a histogram) of the median price change across station-quarter observations. This process can be designed as a series of either supervised or unsupervised machine-learning tasks.

8. Once the classification based on both cyclicality and asymmetry is complete, compute the mean margin and other statistics for each type of observation (as in 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 data set 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 data set (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 choice in terms of cycle-detection accuracy, other methods (including the existing ones) have important roles to play, both as tools for initial inspection and as summary statistics for refinement.

9. Conclusion

We propose scalable methods to detect Edgeworth cycles so that the growing amount of data on fuel prices can be scrutinized. The failure of the existing methods with noisy data suggests further investigation would benefit from distinguishing cyclicality from asymmetry. 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 dis-
cover 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.

References

Assad, Stephanie, Robert Clark, Daniel Ershov, and Lei Xu. 2024. Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. Journal of Political Economy 132:723–71.

Bloch, Harry, and Nick Wills-Johnson. 2010. The Shape and Frequency of Edgeworth Price Cycles in an Australian Retail Gasoline Market. Working Paper No. 201005. Curtin University of Technology, Centre for Research in Applied Economics, Perth.

Bundeskartellamt. 2011. Fuel Sector Inquiry: Final Report in Accordance with §32e GWB. May. Bonn: Bundeskartellamt.

Byrne, David P. 2019. Gasoline Pricing in the Country and the City. Review of Industrial Organization 55:209–35.

Byrne, David P., and Nicolas de Roos. 2019. Learning to Coordinate: A Study in Retail Gasoline. American Economic Review 109:591–619.

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 51:564–77.

Castanias, Rick, and Herb Johnson. 1993. Gas Wars: Retail Gasoline Price Fluctuations. Review of Economics and Statistics 75:171–74.

Chandra, Ambarish, and Mariano Tappata. 2011. Consumer Search and Dynamic Price Dispersion: An Application to Gasoline Markets. RAND Journal of Economics 42:681–704.

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 62:191–228.

Deltas, George. 2008. Retail Gasoline Price Dynamics and Local Market Power. Journal of Industrial Economics 56:613–28.

Doyle, Joseph, Erich Muehlegger, and Krislert Samphantharak. 2010. Edgeworth Cycles Revisited. Energy Economics 32:651–60.

Eckert, Andrew. 2002. Retail Price Cycles and Response Asymmetry. Canadian Journal of Economics 35:52–77.

———. 2003. Retail Price Cycles and the Presence of Small Firms. International Journal of Industrial Organization 21:151–70.

Eckert, Andrew, and Heather Eckert. 2014. Regional Patterns in Gasoline Station Rationalization in Canada. Journal of Industry, Competition, and Trade 14:99–122.

Edgeworth, Francis Ysidro. 1925. The Pure Theory of Monopoly. Pp. 1:111–42 in Papers Relating to Political Economy. London: MacMillan.

Foros, Øystein, and Frode Steen. 2013. Vertical Control and Price Cycles in Gasoline Retailing. Scandinavian Journal of Economics 115:640–61.

Greff, Klaus, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2017. LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems 28:2222–32.

Hansen, Bruce E. 2022. Econometrics. Princeton, NJ: Princeton University Press.
Haucap, Justus, Ulrich Heimeshoff, and Manuel Siekmann. 2017. Fuel Prices and Station Heterogeneity on Retail Gasoline Markets. *Energy Journal* 38:81–103.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9:1735–80.

Ivaldi, Marc, Bruno Jullien, Patrick Rey, Paul Seabright, and Jean Tirole. 2003. *The Economics of Tacit Collusion*. Final Report for the Directorate-General for Competition. Brussels: European Commission.

Klein, Timo. 2021. Autonomous Algorithmic Collusion: Q-Learning under Sequential Pricing. *RAND Journal of Economics* 52:538–58.

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.

———. 2012. Price Leadership and Coordination in Retail Gasoline Markets with Price Cycles. *International Journal of Industrial Organization* 30:342–51.

Lewis, Matthew, and Michael Noel. 2011. The Speed of Gasoline Price Response in Markets with and without Edgeworth Cycles. *Review of Economics and Statistics* 93:672–82.

Linder, Melissa. 2018. Price Cycles in the German Retail Gasoline Market: Competition or Collusion? *Economics Bulletin* 38:593–602.

Lomb, N. R. 1976. Least-Squares Frequency Analysis of Unequally Spaced Data. *Astrophysics and Space Science* 39:447–62.

Martin, Simon. 2018. Market Transparency and Consumer Search: Evidence from the German Retail Gasoline Market. Discussion Paper No. 350. Dusseldorf Institute for Competition Economics, Dusseldorf.

Maskin, Eric, and Jean Tirole. 1988. A Theory of Dynamic Oligopoly, II: Price Competition, Kinked Demand Curves, and Edgeworth Cycles. *Econometrica* 56:571–99.

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

Musolff, Leon. 2021. Algorithmic Pricing Facilitates Tacit Collusion: Evidence from E-commerce. Working paper. Microsoft Research, Redmond, WA.

Noel, Michael D. 2007. Edgeworth Price Cycles: Evidence from the Toronto Retail Gasoline Market. *Journal of Industrial Economics* 55:69–92.

———. 2015. Do Edgeworth Price Cycles Lead to Higher or Lower Prices? *International Journal of Industrial Organization* 42:81–93.

———. 2019. Calendar Synchronization of Gasoline Price Increases. *Journal of Economics and Management Strategy* 28:355–70.

Scargle, Jeffrey D. 1982. Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data. *Astrophysical Journal* 263:835–53.

Siekmann, Manuel. 2017. Characteristics, Causes, and Price Effects: Empirical Evidence of Intraday Edgeworth Cycles. Discussion Paper No. 252. Dusseldorf Institute for Competition Economics, Dusseldorf.

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

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.