FEDERAL RESERVE BANK OF SAN FRANCISCO

WORKING PAPER SERIES

Is China Fudging Its GDP Figures?
Evidence from Trading Partner Data

John G. Fernald, Eric Hsu, and Mark M. Spiegel
Federal Reserve Bank of San Francisco

August 2019

Working Paper 2019-19

https://www.frbsf.org/economic-research/publications/working-papers/2019/19/

Suggested citation:

Fernald, John G., Eric Hsu, Mark M. Spiegel. “Is China Fudging Its GDP Figures? Evidence from Trading Partner Data,” Federal Reserve Bank of San Francisco Working Paper 2019-19. https://doi.org/10.24148/wp2019-19

The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of San Francisco or the Board of Governors of the Federal Reserve System. This paper was produced under the auspices of the Center for Pacific Basin Studies within the Economic Research Department of the Federal Reserve Bank of San Francisco.

Is China Fudging its GDP Figures?
Evidence from Trading Partner Data

John G. Fernald, Eric Hsu, and Mark M. Spiegel[*]
Federal Reserve Bank of San Francisco

August 19, 2019

Abstract

We propose using imports, measured as reported exports of trading partners, as an alternative benchmark to gauge the accuracy of alternative Chinese indicators (including GDP) of fluctuations in economic activity. Externally-reported imports are likely to be relatively well-measured, as well as free from domestic manipulation. Using principal components, we derive activity indices from a wide range of indicators and examine their fit to (trading-partner reported) imports. We choose a preferred index of eight non-GDP indicators (which we call the China Cyclical Activity Tracker, or C-CAT). Comparison with that index and others indicate that Chinese statistics have broadly become more reliable in measuring cyclical fluctuations over time. However, GDP adds little information relative to combinations of other indicators. Moreover, since 2013, Chinese GDP growth has shown little volatility around a gradually slowing trend. Other measures, including the C-CAT and imports, do not show this reduction in volatility. Since 2017, the C-CAT slowed from well above trend to close to trend. As of mid-2019, it was giving the same cyclical signal as GDP.

Keywords: China, GDP, principal components, structural break, forecasting

J.E.L. Classification numbers: C53, C82, E20, F17

[*] We thank Neil Gerstein, Israel Malkin, Andrew Tai, and Bing Wang for excellent research assistance. Japer Hoek, Robert Inklaar, Zheng Liu, Nicolas Petrosky-Nadeau, Zheng Song, Eric Swanson, Harry Wu, and participants in the CUHK/BOFIT conference on China's Financial Intermediation, and the China Conference of International Macroeconomics. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of San Francisco or the Federal Reserve System. E-mails john.fernald@insead.edu; ehsu91@gmail.com; mark.spiegel@sf.frb.org.

1. Introduction

How can we reliably estimate fluctuations in economic activity for a country with statistics of questionable quality? One approach has been to use light as a check on the statistics (Henderson, et al, 2012; Clark et al, 2018). Measured light emissions have considerable high-frequency noise, so this approach serves most naturally as a low-frequency check on statistical quality. But it is often of interest to understand cyclical fluctuations as well. China is a clear example where cyclical fluctuations in economic activity are of first-order interest to many observers, including financial market participants.

In this paper, we propose using imports as a proxy for activity. Imports are one of the best-measured components of GDP and external measures of imports, in the form of exports reported by trading partners, are available. Presumably, these externally-reported statistics are unexposed to domestic manipulation. Moreover, for countries with good statistical systems, we find that imports and measured GDP move closely. But, as would be expected, the co-movement is much weaker for countries with poor statistical systems.

We apply this insight to China. We find that Chinese statistics have, broadly, become more reliable over time in terms of capturing cyclical fluctuations in economic activity. But among possible economic activity indicators that we consider, GDP is merely in the middle of the pack, and its growth rate is excessively smooth since about 2013 relative to other measures of activity. Nevertheless, no single indicator on its own is particularly reliable. Rather, our preferred method for measuring economic activity takes the first principle component of a wide range of indicators such as electricity, industrial production, and rail shipments.

Observers of the Chinese economy have long questioned the accuracy of Chinese output figures.[1] Under any circumstances, measuring Chinese GDP would be difficult given China's rapid growth and undergone extensive structural change (e.g. Holz, 2008). However, many observers also worry that output figures may be distorted, particularly by local and provincial officials in an effort to meet quotas handed down by the government. As a result, many analysts of Chinese economic activity rely instead on alternative, non-GDP indicators.[2]

Skepticism about the accuracy of Chinese data has been shared by prominent Chinese officials. For example, in 2007 current Premier Li Keqiang, was reported as saying that his province's government focused on "alternative indicators," rather than official GDP data (Wikileaks, 2007). Li mentioned three indicators: 1) electricity consumption; 2) the volume of rail cargo, which he suggests is fairly accurately measured because fees are charged for each unit of weight; and 3) the amount of loans disbursed, which may be more accurate because of regulatory oversight. By looking at these three figures, Li said he can measure with relative accuracy the speed of economic growth. Li reportedly said with a smile, "All other figures, especially GDP statistics, are 'for reference only.'"

The challenge in assessing the quality of reported Chinese output figures is to find an independent benchmark to compare with reported aggregate data. One example is Nakamura, et al (2014), who use household consumption data to estimate Engel curves for China. They find that official aggregate consumption data are too smooth relative to levels that would be expected from household spending patterns. More closely related to our paper, Pinovsky and Sala-i-Martin (2016) follow Henderson, et al (2012) and use satellite data on light emissions to gauge growth

[1] See Owyang and Shell (2017) and Sinclair (2012) for extensive references. More recently, Chen, et al (2019) note substantive discrepancies between local and national estimates of industrial output.

[2] For examples of informal press discussions, see Noble (2015), Sharma (2013), and Bradsher (2012).

in economic activity for a cross-section of countries, including China. China's reported GDP growth rate appears to be exceptionally high relative to its growth in observable light. Clark, Pinovsky, and Sala-i-Martin (2018) focus specifically on China. Although the time series on light emissions appears to suffer massive measurement error (e.g., from changes in the sensitivity of satellites over time), they use cross-province variation to assess the informational content of various indicators available regionally. Chen, Chen, Hsieh and Song (2019) use value-added taxes on GDP components as well as local indicators less likely to have been manipulated. They find that GDP growth from 2010-2016 was 1.8 percentage points lower than reported.[3]

We argue that inflation-adjusted imports (measured using trading-partner-reported exports) can serve as a reliable high-frequency measure of fluctuations. Like measured light emissions, these data are reported externally, so they are not subject to manipulation or mismeasurement by Chinese authorities. However, they should be closely associated with economic activity in China, without suffering from the massive measurement error in the light data. Specifically, since the data correspond to Chinese imports, they reflect both the use of intermediate inputs for production—an important aspect of China's economy—as well as finished goods imported for final consumption by Chinese residents. As we show below, although the external sector represents only a portion of economic activity, imports co-move very closely with GDP for many economies.

We take this source of information as an indicator of Chinese economic activity and compare movements in externally-reported exports to China to reported GDP, as well as to various combinations of domestically-reported "alternative indicators" of Chinese activity. If we find that movements in externally-reported exports to China are closely associated with

[3] See also Wu (2014), who estimates that GDP growth from 1977 to 2012 was overstated by 1.8 pp per year.

movements in reported Chinese data, then we conclude that these series are not spurious, but instead are tracking underlying Chinese activity.

Note that this approach tells us about cyclical co-movement, not about the overall level of bias. For example, different series might have different trends for perfectly sensible economic reasons (e.g., structural change in the economy). Indeed, to ensure that the co-movement we detect reflects cyclical fluctuations, we detrend all data prior to estimating the relationships. Hence, our focus is explicitly on uncovering cyclical fluctuations around the trend.

We then turn to the question of the set of indicators that best fits these movements of externally-reported exports. We begin by examining the first principal component of combinations of 14 widely cited and easily available economic indicators, including GDP, produced by Chinese authorities. Our goal is to identify which of these indicators, singly or in combination, best explain China's externally-reported imports.

We begin by considering the performance of each indicator individually. We compare the performances of the first principal component of each indicator, both in-sample and for forecasting out of sample in terms of root-mean squared error (RMSE). We find that electricity usage emerges as our best-fitting individual indicator with estimation conducted over our full sample. Electricity usage also does best both in and out of sample when we repeat the exercise for a sub-sample period beginning after the global financial crisis.

However, while electricity also performs comparably to the first principal component of the Li indicator variables in sample, the Li series does far better out of sample. Similarly, we find that the in-sample fit of the first principal component of all 14 of our indicators combined performs comparably to electricity in-sample, but far outperforms that single indicator out of sample.[4]

In contrast, the link between GDP and externally-reported Chinese imports turns out to be relatively weak. Other individual indicators fit better; and GDP adds little information on activity relative to principal components of many sets of indicators.

Moreover, many of the principal component indices (including the one that includes all 14 indicators) outperforms Li's particular index, both in and out of sample. In particular, although electricity is strongly associated with imports, the other two Li indicators, rail freight and lending, are less important. Nevertheless, we find relatively little sensitivity to the exact group of included activity indicators overall in our comparisons of different groups of predictors.

We find that one of the Li indicators, lending, does particularly poorly as an individual indicator. This finding contrasts not only with Li, but also with Clark et al (2018), who argue that lending is closely related to the cross-provincial pattern of light emissions. In our view, this highlights a potential shortcoming of attempting to use the cross-section on light emissions to identify "good" indicators. Especially in the Great Recession, lending has been used as a countercyclical policy measure to combat growth slowdowns. Hence, lending has a near-zero contemporaneous correlation with GDP, electricity, exports to China, and many other indices of activity that we construct in this paper. But if the endogenous countercyclicality were a response to the aggregate economy, it is likely to be captured by the time fixed effects in Clark et al.'s

[4] Fernald, et al (2013) find that a broad set of activity indicators similarly track the Chinese slowdown from 2010-2012 relatively well, and also outperforms the Li index.

regressions. Hence, lending might be a good measure of relative provincial economic activity in the cross section, without necessarily being a reliable measure of fluctuations in the time series.[5]

Our results do suggest that the accuracy of reported GDP improved during and following the financial crisis, though it subsequently deteriorated again. It becomes far too smooth after 2013 relative to all of our alternative measures of economic activity. We conclude that China's apparent Great Moderation since 2013 is largely spurious.

In the first part of this paper, we look at imports as a measure of economic activity in a cross section of countries. We find that import growth moves closely with GDP for countries with relatively reliable statistical systems.

We then turn to Chinese data, and compare export growth to a wide range of indicators, individually and in combination. For the combinations, we construct the first principal component of all 16,383 possible combinations of these variables and relate them one-by-one to externally reported Chinese imports. Principal components estimation proves useful for yielding a parsimonious specification. Some of the individual indicators that we use might be subject to manipulation or systematic mismeasurement; but, if so, our tests would find that they are not closely related to our externally-reported Chinese-import data. Even if the indicators are informative, they might be noisy. By extracting an activity factor as the first principal component, we reduce the idiosyncratic noise in order to focus on the signal.

This principal-component methodology allows us to focus on a parsimonious relationship and to identify a preferred index of activity. In particular, we relate each combination to externally-reported Chinese imports both in sample (ending in 2013) and out of sample

[5] Lending may be a leading indicator of future activity. We do not consider leads and lags because we would end up with too many potential combinations. Lending could also be important in Clark, et al (2017)'s light-based methodology to the extent this approach is more tailored towards lower frequencies than our use of import data.

(beginning of 2014 through the third quarter of 2018), and then rank each index as a weighted average of in-sample and out-of-sample fit, with weights based on the inverse of the standard deviation of in and out-of-sample RMSEs.

First, we confirm that it is preferable to use a long sample to estimate the factor structure. We reach this conclusion by doing out-of-sample tests of predictive power. In particular, for each of the 16,383 possible combinations of individual indicators, we look at whether the out-of-sample fit is better if the factor structure was estimated over a long sample (starting in 2000) or a short sample (starting in 2008).[6] In 94 percent of cases, the out-of-sample fit is better when the factor structure was estimated on samples that began in 2000. Intuitively, there is a tradeoff between bias (if the factor structure has changed) and precision (if the sample is too short). This finding that factor estimation should be done with a long sample is consistent with the recommendation of Stock and Watson (2016).[7]

We then search for the “best” combination of alternative indicators, including GDP as a potential indicator, based on goodness of fit in and out of sample with our externally-reported Chinese import data. It is not necessarily the case that adding an indicator to an existing set of indicators will improve the fit of the principal component, even in sample. At the same time, a concern with being too parsimonious is that we will select variables that fit well in the specific sample periods we consider. Hence, in cases where two sets of indicators yield identical fit measures, we prefer to choose the large set.

[6] Our out-of-sample period covers 2016Q1-2018Q4. We also examined the robustness of our results to the alternative three-year out-of-sample period 2015Q1-2018Q4.

[7] We also conducted formal Bai and Perron (1998) tests and failed to reject the null of no structural break in the data in favor of either one or two structural break alternatives. These tests are available on request from the authors.

We consider two approaches: We first pursue a “sequential,” approach, adding indicators one at a time from our full group of 14 activity indicators based on their individual fit. This ensures that for a given number of indicators, we have chosen indicators that each fits well individually (and, thus, is a priori plausible). For example, using this approach, the best (average) fit in sample and out-of-sample includes the top six individual indicators. Fit deteriorates somewhat as we add a seventh indicator, and deteriorates a bit more as we continue to add indicators. Second, we re-optimize at each stage completely, choosing the best-fitting combination of each number of indicators without constraining the combinations to include the indicators chosen in the last smaller combination.

This second approach suggest that the best-fitting combination of indicators according to our weighted average of in and out-of-sample fits includes a combination of eight indicators: electricity, exports, industrial production, an index of consumer expectations, fixed asset investment, floor space construction, retail sales, and rail freight. This combination performs best using the unconstrained approach, and so we label it our preferred China cyclical activity tracker (C-CAT).

There are also more parsimonious combinations of six and seven indicators that do almost equally well using our unconstrained approach. One of these, the six-indicator combination, is also the combination that does best under our sequential approach. Relative to our preferred eight-indicator set, it omits retail sales and fixed asset investment. It has an almost identical score both in and out-of-sample, and only does marginally worse in terms of its weighted average score. As we are interested in incorporating as many indicators as possible without sacrificing goodness of fit, we consider the eight-indicator C-CAT our preferred one.

As each of our alternative indicators by construction focuses on specific areas of the China economy, it is plausible that the time series of Chinese imports does not follow those of many of our alternative indicators exactly. But GDP is supposed to be the broadest measure of economic activity. By including GDP as one of the indicators, its variation is included in our measures. However, we find that adding GDP or not adding GDP makes very little differences to the explanatory power of our preferred principal component indices.

Our emerging picture seems to be one where reported GDP is somewhat better at predicting Chinese activity as proxied by externally-reported import data than it used to be—but it is spuriously smooth since 2013. GDP adds at most modestly to the accuracy of the fit of our best combinations of alternative indicators.

It should also be pointed out that our C-CAT should be of use as an alternative measure of Chinese activity, despite recent trade distortions. While the components of our preferred indicator are chosen based on historical fit to Chinese imports, it does not rely on this measure going forward.

In particular, while we only have data through 2019Q2, our analysis does speak to the severity of China's current slowdown. Our preferred C-CAT fell much more than did GDP growth since 2017—but from a well-above trend to a roughly trend pace. By mid-2019, the C-CAT and official GDP gave a similar cyclical signal, as did the all-indicators index. Our results therefore do not indicate that Chinese GDP figures are currently overstating economic activity, at least relative to the degree they did so in the earlier portion of our sample.

The remainder of this paper is divided into six sections: Section 2 discusses the relationship between imports and measured GDP, and how this relationship depends on a country's statistical capacity. Section 3 describes our data and methodology. Section 4 argues for using the full sample to estimate the factor loadings, despite evidence that the quality of statistics has improved over time. Section 5 shows our main results, choosing our preferred set of eight indicators as our "best indicator," And evaluating the recent performance of that indicator relative to reported GDP. Lastly, section 6 concludes.

2. Imports as a measure of activity

The challenge in assessing the reliability of different economic indicators is that we need a benchmark that is highly correlated with true activity but is not, itself, subject to manipulation. In this section, we document that a country's imports fit that bill: Import growth moves closely with GDP growth for countries with relatively reliable statistical systems.

Why would we expect imports to be one of the best measured components of the national accounts? First, the number of importers (and import locations) is typically modest, which makes measurement more manageable. Second, countries have an incentive to measure imports accurately for tariff purposes. Third, data on imports are available from external sources, reported as trading partner's bilateral exports to the country in question.

In countries with less-advanced statistical systems, we would expect the relationship between imports and measured GDP to deteriorate simply because measured GDP becomes less accurate.[8] The reduced accuracy of measured GDP should then reduce its correlation with imports. In contrast, for the reasons noted above (including the external verification), there is little reason to think that the correlation between imports and *true* economic activity deteriorates.

To assess these conjectures, we look at cross-country data to see the relationship with statistical capacity. We use data on 165 countries from the Penn World Tables (release 9.0). For

[8] For example, Subramanian (2019) has argued that official Indian estimates overstate GDP growth between 2011 and 2016 by about 2.5 percentage points.

each country, we calculate the correlation of growth in real imports and real GDP from 1990 to 2014, using national source data (the data that underlie the more-frequently used purchasing-power-parity data). For each country, an earlier version of the PWT (release 6.1) had a judgmental measure of statistical capacity, which ranked the countries from A (highest) to D (lowest).[9]

Table 1 relates the import-GDP correlation across countries to statistical capacity and other control variables. The control variables are country size (GDP in 1990 in international dollars) and initial income per capita (GDP per capita in 1990, in international dollars). Overall GDP could be of either sign. There may be scale economies in data collection that results in greater effort in generating statistics, but larger economies may also be less open and therefore the correlation may be reduced holding statistical quality constant.

Income per capita could, independently, be associated with the correlation between imports and GDP. For example, the structure of the economy—say, goods relative to services—might systematically be related to the level of income. Since many low-income economies have low statistical capacity, we want to be sure that statistical capacity is not simply proxying for income. (We use 1990 values for GDP and GDP per capita, but using average values is very similar).[10]

The first column examines the impact of GDP and GDP per capita. The unconditional correlation is, in fact, positively and significantly related to the initial level of GDP: Larger

[9] See Dawson et al., 2001, for a discussion of how this measure is constructed. They note that for some issues in the growth literature, accounting for statistical capacity is important. Henderson et al (2012) use a different measure of statistical capacity from the World Bank, but that one is *only* available for developing economies, not for the full universe of countries.

[10] We also looked at measures of openness as a control. But we found no relationship with the correlation between imports and GDP, so we do not show this variable in the table.

countries have a higher estimated correlation. In contrast, the correlation is insignificantly related to GDP per capita.

The second column shows that countries with poor statistical capacity (C or D) do indeed have a notably lower correlation between growth in GDP and in imports. The constant term shows that a country with an A-rated statistical capacity (the omitted category in the regression) has a correlation that is nearly 0.8. For these countries, imports move quite closely with GDP. However, for a country with D-rated statistical capacity, the correlation is substantially lower at 0.36 (0.788 plus the coefficient on statistical capacity D of -0.426); the difference is highly statistically significant. We also obtain a statistically significantly lower correlation for a country with a C-rated statistical capacity.

Table 1: Statistical capacity and correlation of growth in imports and GDP

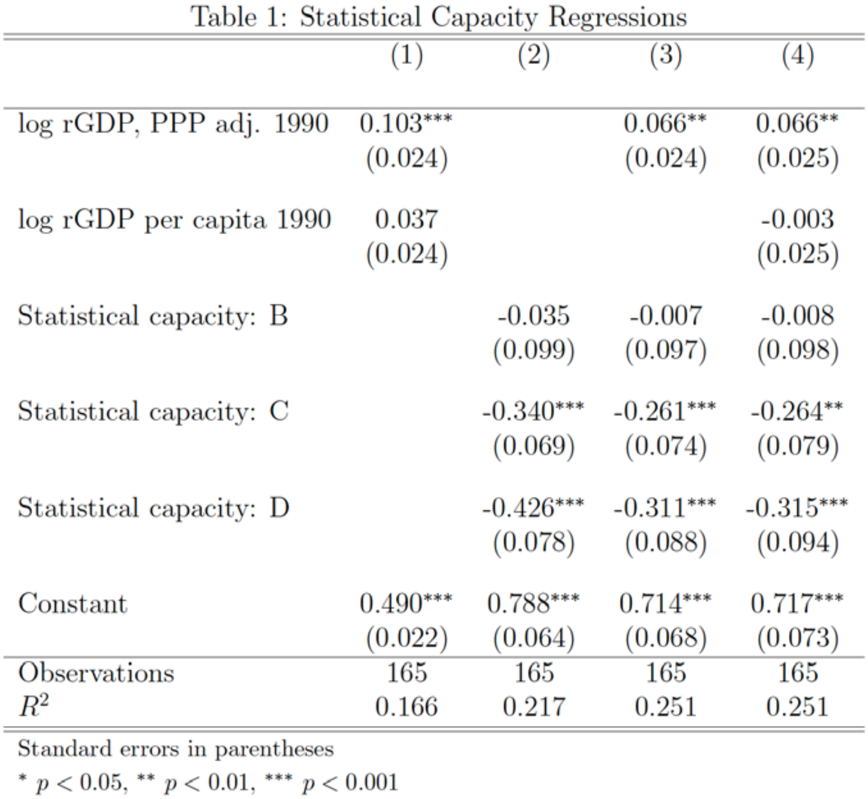


Table 1: Statistical Capacity Regressions

	(1)	(2)	(3)	(4)
log rGDP, PPP adj. 1990	0.103***		0.066**	0.066**
	(0.024)		(0.024)	(0.025)
log rGDP per capita 1990	0.037			-0.003
	(0.024)			(0.025)
Statistical capacity: B		-0.035	-0.007	-0.008
		(0.099)	(0.097)	(0.098)
Statistical capacity: C		-0.340***	-0.261***	-0.264**
		(0.069)	(0.074)	(0.079)
Statistical capacity: D		-0.426***	-0.311***	-0.315***
		(0.078)	(0.088)	(0.094)
Constant	0.490***	0.788***	0.714***	0.717***
	(0.022)	(0.064)	(0.068)	(0.073)
Observations	165	165	165	165
R^2	0.166	0.217	0.251	0.251

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Regressions of import-GDP correlation across countries on statistical capacity and other control variables. Measures of statistical capacity from PWT (release 6.1), which ranked countries from A (highest) to D (lowest). See Dawson, et al (2001) for details. Standard errors in parentheses.

The third column adds back GDP, and the fourth column also adds GDP per capita. The results in these columns only modestly mitigate the effects shown in the first two columns. Thus, the statistical evidence suggests that while in countries with relatively good statistical systems, GDP and imports move closely with each other, that is not the case for countries with limited statistical capacities.

For comparison, the United States has a statistical rating of A and a correlation of above 0.9. China has a statistical rating of D and a correlation of about 0.6. China's correlation is above its expected value, conditional on its statistical capacity rating of D. However, it is below what would be expected for a country with an A or B rating. The next section considers what we can learn about the quality of China's various statistical releases.

We conclude this section by acknowledging that imports are also an imperfect and likely noisy measure of economic activity in China. Moreover, structural change might mean that the relationship between this indicator and other indicators has changed over time. For example, in annual U.S. data for the 30 years (1986-2016), import growth is more highly correlated with growth in goods (0.9) than in structures (0.6) or services (0.5). That said, the correlation is highly significant even with these different components of activity. And, even if it noisy and imperfect, there is little reason to think it is biased.

In our empirical work for China, we consider whether the relationship has changed by looking at time variation. A priori, structural change seems like it could cause relationships to attenuate over time, given that China's service sector has grown in importance relative to the traded sector. However, there is little decline in the share of imports in Chinese GDP over the course of our sample. Empirically, the relationship between GDP and imports has improved over time.

3. Chinese data

Our goal is to use the insight from the previous section about the information content of imports to develop a reliable indicator of activity in China.[11] This section discusses what Chinese data we use to achieve that goal.

[11] In principle, this might point towards just using imports alone as a measure of activity. However, trade volumes might be disproportionately susceptible to policy distortions, for example in the current trade dispute between China and the United States. We therefore would not want to base our estimate of Chinese activity solely on trade measures, although in practice, our analysis below suggests that Chinese exports are an important indicator throughout our sample.

A. *Measuring China's imports*

For any country where the accounts are suspect, including China, there is a question of whether the import statistics themselves are accurate. As noted in the introduction, a key advantage of imports is that they can also be measured using trading-partner exports. For both economic and statistical reasons, we combine exports to China and Hong Kong for these purposes. Economically, many of the goods that are exported to Hong Kong from non-China sources are destined for the Chinese mainland.[12] Statistically, authorities in, say, the United States may plausibly have changed the degree to which they are able to track the ultimate destination over time—that is, a good that previously would have been recorded as an export to Hong Kong might now be recorded as an export to China. Using the combination of Hong Kong and China makes the data more comparable over time.[13]

For our main analysis, we use trading-partner-reported export data, since measurement error in this indicator should be independent of the measurement error in China-source economic indicators. This source of data on China's imports is not controlled in any manner by Chinese authorities. (Henceforth, when we refer to imports, it's always as reported by trading partners.) Trading-partner governments have no apparent incentive to misrepresent their trade volumes with China. Of course, the rapid growth of trade with China could still cause some measurement challenges for these countries. However, these data still have the advantage of being measured

[12] For example, in 2016 over US$400 billion in goods were re-exported through Hong Kong from and to the Mainland (https://www.tid.gov.hk/english/aboutus/publications/factsheet/china.html).

[13] Fernald, Edison, and Loungani (1999) argue for combining Hong Kong with China. We confirm in the data appendix that imports by China and Hong Kong imports (henceforth referred to as "China's imports") move very closely with its trading-partner-reported exports (see Appendix Figure A2). The trading-partner data line up better for China plus Hong Kong than for China alone.

at foreign ports. Moreover, while Chinese trade is growing as a share of total trade for these countries, overall trade is not growing nearly so fast. So tracking trade volumes, including those destined for or originating from China, is arguably less challenging.

Using the IMF's Direction of Trade Statistics (DOTS), it is straightforward to measure world trading-partner exports to China and Hong Kong (excluding exports from China to Hong Kong and vice versa). We obtain similar results to those reported later in this paper when we use narrower sets of countries—such as exports from the United States, the Euro Area, and Japan. Because imports represent intermediate inputs and final consumption or investment goods, they are likely to be correlated with overall activity. To calculate real imports, we deflate with a China-specific export deflator, described in the appendix.[14]

B. *Individual data series*

From Chinese-source data, we identified 14 potential activity indicators on the basis of data availability and a priori plausibility—GDP plus 13 non-GDP variables. The 14 indicators are all available from the beginning of our sample (the end of 2000), and were downloaded from CEIC Asia. Examples include electricity use, industrial production, rail freight, and new property construction. The full list of indicators are described in the data appendix and also listed in the tables in the next section. Although GDP is of independent interest, for our main purpose ("what is the best index of activity in China") we consider GDP as just one of a list of possible indicators to examine.

[14] DOTS measures trade in U.S. dollars (converted with market exchange rates), so we need a dollar-based deflator. We use U.S. product-level export prices, weighted by the product-level imports of China and Hong Kong. U.S. measures of prices are considered relatively reliable; and any biases are likely to be unrelated to economic conditions in China. Even if the weights were not reliable (though there is little reason for errors in import composition), the bias for the overall deflator is likely to be small.

More than 13 non-GDP indicators are available for the full sample (e.g., the inward flow of FDI). However, these were a priori less obviously linked to Chinese economic activity. Furthermore, in preliminary analysis, we found little statistical relationship with imports or other contemporaneous measures of economic activity. Finally, our selection method, discussed below, becomes computationally unwieldly as the number of indicators grows.

To control for seasonal factors, we use all variables in four-quarter changes.[15] In principle, we could use the Census X-12 program to control for seasonality. For our purposes, we prefer the simple and transparent year-over-year change. In addition, Wright (2017) raises questions about the reliability of the X-12 procedure.

Before doing any statistical analysis, we follow Stock and Watson (2016) and detrend all individual indicators with a biweight filter. The biweight filter is essentially a smooth two-sided filter that becomes increasingly one-sided at the end points. The reason for filtering is that individual series have different trends—which can be misleading, since our principal component indices will attempt to fit those trends as well as the fluctuations.[16]

For example, without filtering, Chinese GDP growth in 2019Q2 (6.2 percent year-over-year) was worse than at the trough of the Great Recession. Yet, while growth was relatively slow, that appeared to be a trend development, rather than as an indication of slow cyclical growth. The appendix shows the raw individual indicators, their estimated trends, and the detrended indicators (Figure A1). In all cases, the detrended data have been normalized to have mean zero and unit standard deviation.

[15] Many of the series are available monthly, but we convert all data to quarterly terms. Doing so facilitates comparisons with quarterly GDP data, smooths some high-frequency measurement error, and avoids problems with the timing of the Chinese New Year (which sometimes occurs in January, sometimes in February, and sometimes overlaps both).

[16] The biweight filter suggests a downward current trend in Chinese growth. We concentrate here on tracking cyclical movements.

A question is what bandwidth to use for the filter. For U.S. data, Fernald, Hall, Stock, and Watson (2017) use a biweight filter with bandwidth of 60 quarters before estimating a factor model. That filter yields relatively smooth trends, which are not too sensitive to end-point issues. Even there, however, it is clear that the filter is too smooth to capture the trend for some U.S. series, such as total factor productivity.

For China, the changes in trend growth since 2000 are much sharper than for the United States, and a more responsive filter appears to fit the data better. For this reason, we use a filtering parameter of 24 quarters, which is flexible enough to fit the trends reasonably well. Despite this flexibility, the trend does not appear overly sensitive to end points, so revisions to real-time estimates of the trend are not too large. For example, we estimated the trend through 2009—when the cyclical deviation from any trend was clearly very large—and compared the real-time trend to the revised trend. While there were revisions, they were not extreme. The main results do not appear driven by the choice of filtering.

C. Principal components

To identify a "preferred" index of China's cyclical fluctuations, we focus on principal component indices from potential sets of individual indicators. The reason is that indicators that move closely with China's imports in any given sample sometimes fit less well out of sample. Using the first principal component from a set of indicators helps minimize this problem.

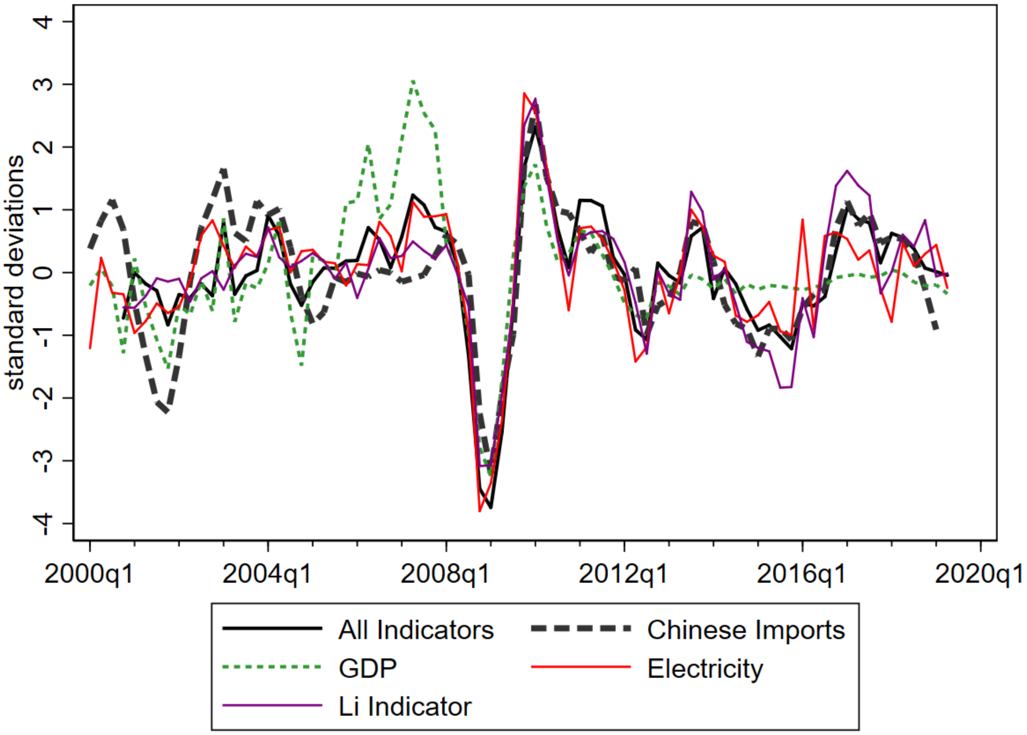
For example, an extremely misleading approach to using the individual indicators would arise if one simply regressed China's imports on all 14 of the indicators plus GDP. Such a regression has a high R^2 even though, because of multicollinearity, few of these indicators are statistically significant. However, because of overfitting, this approach performs very poorly out of sample relative to a more parsimonious specification.[17]

Principal components help minimize the risk of spurious fit by capturing the key common information in the indicators — known as “activity factors” — in a parsimonious way. Principal components are defined by the property that all factors (or components) are orthogonal, with the first component explaining the maximum variation in the included data, the second one explaining the second most variation, and so forth.

One extremely parsimonious set of indicators is the index that includes GDP alone—a single indicator. The a priori justification is that GDP is, in principle, the broadest measure of economic activity. At the other extreme, an a priori reasonable benchmark index is the first principal component of *all* of our individual indicators, including GDP. That benchmark is agnostic about which indicators are informative or uninformative, and whether that informational content has changed over time.

[17] For example, we regressed import growth on all 14 indicators as separate right-hand-side variables from the start of our sample until end-2015 and predicted out-of sample thereafter. For comparison, we also regressed import growth on the first principal component of these indicators, as well as the first principal component of the three Li indicators. As expected, the regression with all 14 indicators individually had the lowest (best) RMSE in sample, 0.34 versus 0.52 for the first principal component of the 14 indicators and and 0.65 for the first principal component of the Li indicators respectively. However, the regression with all 14 indicators included had a higher RMSE out of sample than the first principal component (0.57 versus 0.49). The out-of-sample results for the Li indicators were modestly worse, with an RMSE of 0.58.

Figure 1 shows selected indicators along with real exports to China. All variables represent year-over-year growth rates and are normalized to have mean zero and unit standard deviation. The indicators shown are electricity, which is often taken as a proxy measure for activity in China, and is our best-fitting individual activity indicator below; the first principal component of all 14 indicators ("all indicators"); the "Li" combination of three indicators (the first principal component of electricity, bank lending, and rail cargo); and GDP. Figure 1: Indicators of economic activity in China



Note: All series represent four-quarter growth rates in economic activity relative to trend, and are normalized to be mean zero and with a unit standard deviation. China imports are measured as real trading-partner exports to China and Hong Kong. "All indicators" is the first principal component of all 14 individual indicators (including GDP). "Li" is the first principal component of electricity, lending, and rail cargo. All series are measured 2000Q4-2019Q2 except for imports, which runs through 2018Q4. See text for further details.

Clearly, the all-indicators activity factor and imports are very highly correlated. For example, during the global financial crisis, both series drop about 3 standard deviations below their respective means. In the recovery, both series rise to above 2 standard deviations above their means. Thus, reassuringly, imports and the activity factor tell the same story about economic activity. The Li indicator is a modestly poorer fit, but in general tends to come close to both imports and the all-indicators activity factor. However, the relationship of reported GDP with either the activity factor or imports is less strong. The correlation is still positive and significant, but GDP rises more prior to the crisis than either imports or the activity factor prior to the global financial crisis.

A key goal of the sections that follow is to identify which indicators (including GDP), or combinations thereof, are particularly informative, in terms of correlation with our externally-reported import indicator. It is possible that a more parsimonious index will be an even better index of economic activity. This could happen either if an additional variable was biased (e.g., because of political pressure), or alternatively, if it was largely idiosyncratic—unrelated to imports and overall economic activity. As a result, the first principal component of a larger set of indicators might be less accurate as a measure of activity because it is trying to explain that biased or idiosyncratic variation as well as the systematic "true" variation in overall economic activity.

We proceed by constructing the first principal component of all possible subsets, other than the null set, of these 14 variables (GDP plus the 13 non-GDP indicators), considering a total of 16,383 combinations. For example, 14 of the combinations have just a single indicator (each of the 14 variables); at the other extreme, one combination uses all 14 variables at the same time (our "all indicators" factor plotted in Figure 1).

For each subset, we then regress growth in Chinese imports from China's top ten trading partners on the first principal component as well as exchange rate values (which plausibly affect import levels independently of output). Our baseline specification is thus

$$\Delta^4 m_t = c + \beta_i PC_{1t} + \gamma \Delta^4 RMB_t + \eta_t \qquad (1)$$

where $\Delta^4 m_t$ is reported quarterly growth in real Chinese imports from (measured as real exports to China by) the United States, the euro area, and Japan; PC_{1t} is the contemporaneous value of the first principal component from the year-over-year growth in the chosen set of alternative indicators of Chinese economic activity; $\Delta^4 RMB_t$ is the four-quarter change in the renminbi-dollar exchange rate; and η_t is an error term. We estimate with ordinary least squares and show Newey-West standard errors that allow for heteroskedasticity and autocorrelation.

The reason for controlling for the exchange rate is to control for non-activity factors that might affect imports. Conceptually, demand for imports in China could depend on two factors. The first is direct final demand for consumption, investment, or government. The second is imported intermediates that are destined for re-export after some further processing. Some but not all of the activity included in the second category reflects activity being done in China. But some reflects derived demand from activity taking place in countries to which China exports.

Now suppose that the RMB depreciates. That might directly lead to a shift of Chinese domestic demand towards domestically produced goods and away from the now-more-expensive foreign goods. It also makes imported intermediates more costly, which might affect the incentives to use those parts rather than domestically produced ones. By controlling for the exchange rate, we allow the regression to control for these non-activity channels.

Of course, in-sample (IS) and out-of-sample (OS) fits might be quite different. This is most obvious for the combinations that comprise only a single indicator. Purely by chance, some indicator might happen to move closely with China’s imports during a given sample. To guard against this concern, we seek combinations of indicators that are not only a priori plausible but that perform well both in-sample and out-of-sample.

4. Initial Analysis: What sample to use, given possible structural change?

China's economy has changed dramatically in recent decades. That might point towards wanting to focus on a relatively recent time period. However, the literature on forecasting in the presence of structural change (e.g., Pesaran, Pick, and Pranovich, 2013) finds that the loss of estimated precision from using a shorter sample period can be more important than the bias in the true coefficient caused by structural change. In this section, we find that it is preferable to use a relatively long sample period to estimate the econometric relationships.

Conceptually, the relationship in equation (1) could change in two ways that would affect the reliability of the relationship. First, the coefficient β_i on some particular principal component index (PC_i) could change over time, perhaps reflecting structural change. Second, even if the β_i do not change, the statistical fit of the relationship could change if the variance of η_t changes. (η_t could, of course, reflect idiosyncratic noise in imports or in the principal components.) In the second case, an indicator might be unreliable even if it is an unbiased estimate of activity.

First, Bai-Perron (2003) break tests find no evidence of structural changes in the parameters β_i for two a priori sensible principal-component indices (the PC_i's in equation (1)). Specifically, in results not shown, we estimated equation (1) with the all-indicators index and the Li index (the principal component of electricity, bank lending, and rail cargo).[18] With (filtered) import growth as the left-hand-side variable, as in equation (1), there is no evidence of instability in the regression coefficients. (The tests do find instability in the relationship when GDP growth is used as the PC_i measure, as we discuss below.)

Second, to look at the overall evidence of change in the statistical relationship, we look at the adjusted R^2 from 24-quarter rolling regressions of China's imports on selected activity

[18] These results are available on request from the authors.

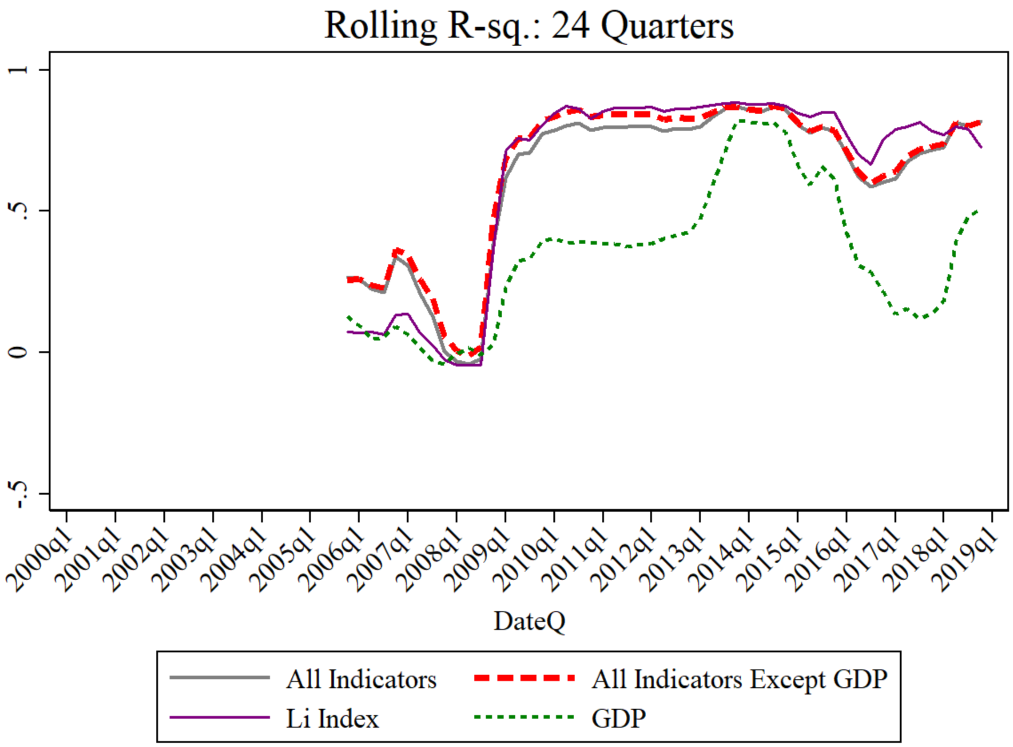
indicators. (For these purposes, we drop the exchange rate from equation (1) to ensure that changes in fit over time reflect the activity factor, not the exchange rate.) The R^2 shows the ratio of "explained" variance $\hat{\beta}_i PC_{1t}$ to total variance; correspondingly, it rises if the idiosyncratic variance η falls relative to the explained variance of China's imports. In contrast, the break tests look more narrowly for subsample changes in β. A good indicator not only has a stable relationship with activity, as the break tests indicate, but also explains a lot of the variation in imports (high R^2).

Figure 2 shows the rolling R^2 estimates. All of the indices, especially GDP, fit poorly prior to 2008, but fit better thereafter. Some of this is just the sharp decline and rebound during the Great Recession itself that was common to imports and all of the activity indices. However, even in samples that end 2016 or after (so they are not affected by the sharp downturn of the Great Recession), the indices fit better than before 2008. Other than GDP, the indices fit *much* better than before 2008.

For GDP, the relationship deteriorates again towards the end of the sample before improving a bit at the very end. Even that improvement, however, is somewhat misleading. Looking back at Figure 1, GDP appears extremely smooth in the past few years relative to other indicators. However, if one looks closely, the wiggles are somewhat correlated, which is what the regression picks up. That is, the direction of the change in GDP growth is accurate, even if the magnitude of the change in growth is understated.[19]

[19] This is picked up by the break tests, which find that the coefficient on GDP becomes much larger, statistically and economically, after 2013.

Figure 2: Rolling (Adjusted) R-Squareds



Note: Rolling adjusted R^{2}s from regressing China's import growth on the activity indicator shown; regressions are run over the 24 quarters ending at the date indicated on the horizontal axis. See text for further details.

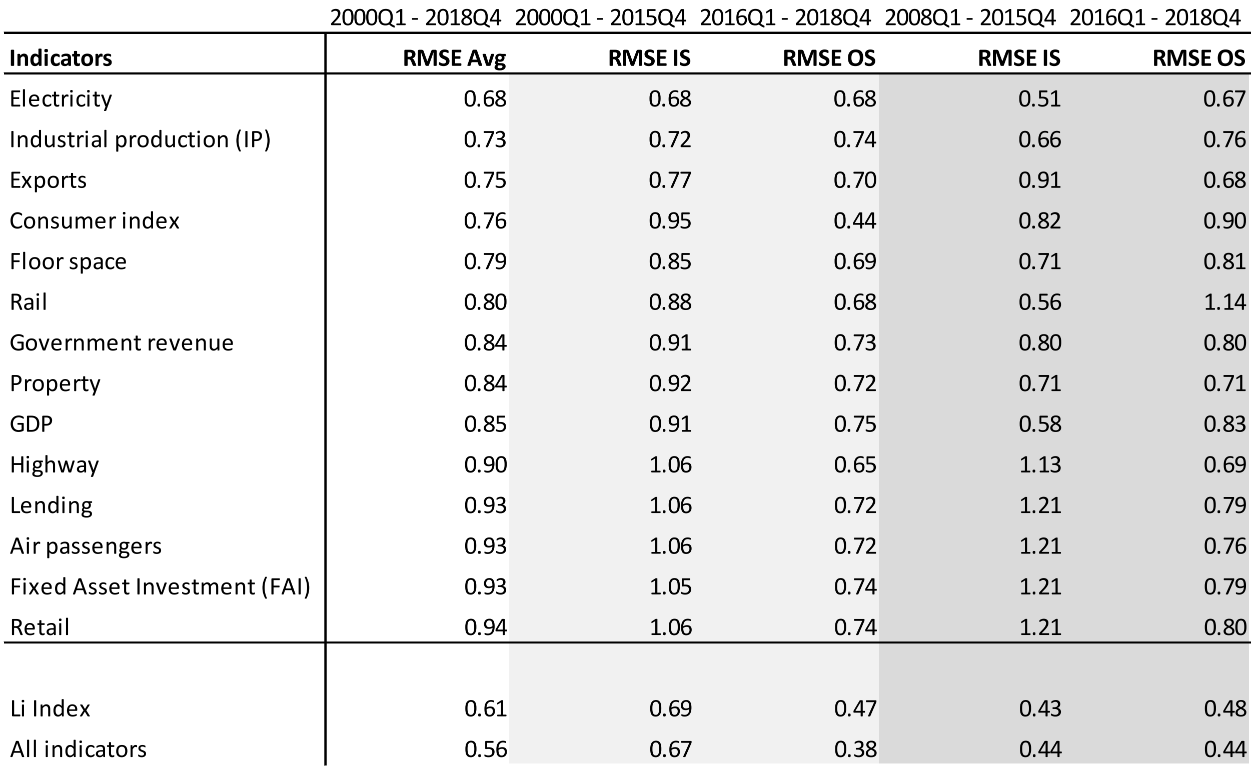
Note that the all-indicators combination that excludes GDP has equal, or modestly better, performance throughout relative to the all-indicators combination that includes GDP. At the end of our sample, even as GDP on its own deteriorates somewhat, including GDP with the other indicators makes little difference.

In light of the break tests and rolling regression R^2's, it is unclear what sample to use. The break tests do not suggest an econometric reason for shortening the sample, but the R^2's suggest that the relationship is much closer after 2008.

To address this uncertainty, we perform out-of-sample exercises after estimating the relationship in equation (1) for different in-sample periods. We find that using the full sample, starting in 2000, unambiguously does better for explaining exports to China out of sample than starting in 2008. Specifically, for each of the 16,383 combinations of indicators that we consider, we estimated equation (1) for the 2000-2015 period and then the 2008-2015 period. Then we looked at how well the estimated relationships fit in the 12-quarters of 2016 to 2018. In only 6.9 percent of the combinations did the out-of-sample (2016-2018) estimates have a lower RMSE when we used the 2008-2015 period for estimation rather than the 2000-15 period. In other words, in the vast majority of cases, we do better by using the full sample for estimating the relationship between exports to China and indicators.

Table 2 shows a subset of these results. Specifically, it shows the in-sample and out-of-sample fit for the 14 individual indicators we use plus several combinations. The in-sample root mean squared errors (RMSEs) are not comparable, because they correspond to different sample periods. But the out-of-sample RMSEs all correspond to the 2016Q1-18Q4 period; they differ only because the coefficients were estimated over different periods. For 11 out of the 14 indicators, the out-of-sample RMSE is lower (i.e., better) when the relationship is estimated over the longer 2000-2015 sample than over the shorter 2008-2015 sample. We also obtain lower RMSE’s out-of-sample for the Li index and the all-indicators index when estimated over the longer time period.

Table 2: Out-of-sample results for different in-sample periods



	2000Q1 - 2018Q4	2000Q1 - 2015Q4	2016Q1 - 2018Q4	2008Q1 - 2015Q4	2016Q1 - 2018Q4
Indicators	**RMSE Avg**	**RMSE IS**	**RMSE OS**	**RMSE IS**	**RMSE OS**
Electricity	0.68	0.68	0.68	0.51	0.67
Industrial production (IP)	0.73	0.72	0.74	0.66	0.76
Exports	0.75	0.77	0.70	0.91	0.68
Consumer index	0.76	0.95	0.44	0.82	0.90
Floor space	0.79	0.85	0.69	0.71	0.81
Rail	0.80	0.88	0.68	0.56	1.14
Government revenue	0.84	0.91	0.73	0.80	0.80
Property	0.84	0.92	0.72	0.71	0.71
GDP	0.85	0.91	0.75	0.58	0.83
Highway	0.90	1.06	0.65	1.13	0.69
Lending	0.93	1.06	0.72	1.21	0.79
Air passengers	0.93	1.06	0.72	1.21	0.76
Fixed Asset Investment (FAI)	0.93	1.05	0.74	1.21	0.79
Retail	0.94	1.06	0.74	1.21	0.80
Li Index	0.61	0.69	0.47	0.43	0.48
All indicators	0.56	0.67	0.38	0.44	0.44

Note: Indicators are ordered based on increasing RMSE average in the first column. This RMSE average is the weighted average of the in-sample (IS) and out-of-sample (OS, 2016Q1-2018Q4) RMSEs, where the in-sample period is 2000Q4-2015Q4. The weights are inverses of the cross-sectional standard deviations (across all 16,383 combinations of indictors) of the RMSEs in the IS and OS periods.

Before moving on, we note that in Table 2, the first nine individual indicators (through GDP) all have in-sample RMSEs that are less than one. Because the import index on the left-hand-side is normalized to have a unit standard deviation, the regression with no explanatory variables would have an RMSE of one. Thus, the first nine variables reduce the RMSE relative to omitting the variable, whereas the bottom five actually do worse. The divide is particularly sharp with the 2008-2014 in-sample period. We also note that while there is some tendency for indicators that do well in-sample to also do well out-of-sample, with low RMSEs, the ranking is far from one-to-one.[20]

In sum, the break tests and the in-sample/out-of-sample tests both suggest using a long sample. We follow that approach in the next section. The rolling estimates do find that even GDP is more informative after 2008. That said, these do not suggest that one should focus solely on GDP. Rather, they suggest that it is still preferable to use a principal component of a wide range of indicators. In the next section, we consider whether we should use a parsimonious set of indicators to form our principal component. `

5. Relative performances of individual and combinations of activity indicators

A. Results for individual indices and sequential activity indicator selection

Given the superior performance of estimates fitted over our full sample, we next investigate the relative performances of a variety of combinations of activity indicators over the full sample. We begin by summarizing the individual performances of our activity indicators. The estimated parameter values are not interesting per se, so we do not show them. We instead focus on (i) index names and sets; and (ii) fit as measured by RMSE.

We are interested in both in-sample performance and out-of-sample prediction. As we demonstrate below, the relative quality of fit of combinations of activity indicators in and out-of-sample can differ markedly, and it seems plausible that the structural changes that have recently taken place in the Chinese economy may lead some indicators to erroneously indicate strong or weak performances out of sample.

[20] When we average the IS and OS RMSEs, we weight by inverse cross-sectional standard deviations in RMSEs across the 16,383 combinations (scaled by the sum of these weights, so that the weights sum to one). The standard deviation is higher in the in-sample period than the out-of-sample period.

It is unclear how to weight the relative importance of in- and out-of-sample fit. In response, we characterize performance in terms of the weighted average RMSE in and out of sample, where weights are measured as the inverse of the variance of the population of in and out-of-sample RMSEs respectively. There is also no consensus about the share of a time series sample that should be allotted to the estimation of weights on activity indicators in sample and that allotted to gauging the performances of these indicator combinations out of sample [e.g. Hansen and Timmermann (2012)].

We first consider the relative performances of indicator combinations by pursuing a sequential approach, based on the indicators' individual performances. From Table 2 (left two columns of numbers), we start by identifying the best-performing single indicator in terms of weighted average RMSE. Next, we add the second-best individual indicator to form an index of the first two individual activity indicators, and so on. The rationale for this approach is that we are sequentially adding variables that, individually, have explanatory power. So the resulting indices include only variables that have a priori justification.

Our results are shown in Table 3. Our best-performing individual indicator is electricity, which outperforms all other individual indicators with RMSEs both in and out of sample of 0.68. We thus begin with electricity as our single-indicator combination. As shown in Table 3, we get a modest improvement in average performance by adding the second activity indicator, "Industrial production," as the weighted average RMSE drops from 0.68 to 0.66. We get a modest improvement in both in-sample RMSE, which drops from 0.68 to 0.65, and in out-of-sample RMSE, which drops from 0.68 to 0.66.

Table 3: Sequential indicators and average IS-OS rank

			2000Q4-2015Q4	2016Q1-2018Q4
NumVars	**Variables**	**RMSE Avg**	**RMSE IS**	**RMSE OS**
1	Electricity	0.68	0.68	0.68
2	Electricity IP	0.66	0.65	0.66
3	Electricity IP Exports	0.60	0.57	0.63
4	Electricity IP Exports ConsumerIndex	0.57	0.63	0.47
5	Electricity IP Exports ConsumerIndex FloorSpace	0.56	0.62	0.46
6	Electricity IP Exports ConsumerIndex FloorSpace Rail	0.50	0.62	0.31
7	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev	0.51	0.63	0.31
8	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev Property	0.52	0.65	0.32
9	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev Property GDP	0.56	0.67	0.37
10	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev Property GDP Highway	0.56	0.68	0.37
11	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev Property GDP Highway Lending	0.56	0.67	0.37
12	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev Property GDP Highway Lending AirPassengers	0.56	0.67	0.38
13	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev Property GDP Highway Lending AirPassengers FAI	0.56	0.67	0.39
14	Electricity IP Exports ConsumerIndex FloorSpace Rail GovtRev Property GDP Highway Lending AirPassengers FAI Retail	0.56	0.67	0.38

Note: RMSEs in sample, out-of-sample, and (weighted) RMSE averages from regressing import growth on the first principal component of the list of indicators shown. (All regressions also include the change in the real exchange rate). Indicators are sequentially appended to the list according to their individual ranking from Table 2. See also notes to Table 2.

We obtain larger improvement in average RMSE by going to three indicators with the addition of exports. This drops the average RMSE to 0.60. The improvement is primarily driven by a drop in in-sample RMSE from 0.60 to 0.57, but the out-of-sample RMSE also falls from 0.66 to 0.63.

Adding three more indicators, the consumer index, floor space, and rail freight, gives us our best-fitting combination of six activity indicators under this sequential method. Using this combination of activity indicators, the in-sample RMSE falls to 0.62, while the out-of-sample RMSE falls to 0.31. Overall, the weighted average RMSE falls to 0.50.

However, adding additional indicators under this method fails to improve the fit. For example, the combination with the addition of the seventh individual indicator, government revenue, had a modestly larger average RMSE than our best-performing 6-indicator combination (0.51). Overall, it can be seen that average RMSEs are not monotonically declining with the addition of more activity indicators under the sequential method.

There are a number of notable patterns to the sequentially chosen sets of indicators. First, our best set of 6 indicators is not much better than those with larger numbers of indicators, including the all indicators set, which has a weighted-average RMSE of 0.56. Still, there are some apparent gains from parsimony, at least given the in-sample and out-of-sample periods considered.

Second, the indicators differ somewhat in their relative in and out of sample performances. Our best-performing combinations of indicators in-sample are the 5 and 6-variable combinations, the former of which includes electricity, industrial production, exports, the consumer index, and floor space while the latter adds rail freight. In contrast, the best-performing out-of-sample combinations are the 6 and 7-variable combinations, the latter of which adds the activity indicator government revenue.

Finally, note that the introduction of the GDP indicator in the 9-variable activity indicator combination raises average RMSE relative to our 8-variable combination. This is consistent with our finding above that among the individual indicators, GDP is somewhat below average (indeed, out of sample it is the worst of the individual indicators). As such, we find that GDP is informative, but that we can do a much better job of predicting Chinese economic activity when we combine GDP with other informative indicators.

B. Unrestricted combinations of activity indicators

The sequential method is computationally simple, and it ensures that we end up with an index where indicators individually have explanatory power. However, it does not necessarily yield the best combination of activity indicators. For example, the top indicators could all contain essentially the same information on activity (say, on manufacturing production), whereas a lower-ranked individual indicator might contain independent information on activity (say, on services). To obtain those combinations, we examine all possible combinations of the individual activity indicators for each number of potential indicators from one to fourteen. We then again choose the set of indicators for each number with minimum weighted average in-sample and out-of-sample RMSE.

Our results are shown in Table 4. We continue to observe wide discrepancies in relative in and out-of-sample performances for the various activity indicator combinations. Tautologically, electricity remains our best-performing individual indicator. Our next factor adds exports, which lowers our weighted average RMSE from 0.68 to 0.60.

Our best combination of three indicators includes electricity, exports and rail. This combination exhibits 0.61 RMSE in sample, but achieves a 0.43 RMSE out of sample for a weighted RMSE average of 0.54.

Table 4: Average IS-OS Rank

			2000Q4-2015Q4	2016Q1-2018Q4
NumVars	**Variables**	**RMSE Avg**	**RMSE IS**	**RMSE OS**
1	Electricity	0.68	0.68	0.68
2	Electricity Exports	0.60	0.58	0.62
3	Electricity Exports Rail	0.54	0.61	0.43
4	Electricity Exports IP Rail	0.51	0.56	0.42
5	Electricity Exports IP Lending Rail	0.50	0.56	0.42
6	ConsumerIndex Electricity Exports FloorSpace IP Rail	0.50	0.62	0.31
7	ConsumerIndex Electricity Exports FAI FloorSpace IP Rail	0.50	0.62	0.31
8	ConsumerIndex Electricity Exports FAI FloorSpace IP Rail Retail	0.50	0.62	0.31
9	ConsumerIndex Electricity Exports FAI FloorSpace IP Lending Rail Retail	0.50	0.61	0.33
10	ConsumerIndex Electricity Exports FAI FloorSpace GovtRev IP Lending Rail Retail	0.51	0.62	0.32
11	AirPassengers ConsumerIndex Electricity Exports FAI FloorSpace GovtRev IP Lending Rail Retail	0.52	0.63	0.35
12	AirPassengers ConsumerIndex Electricity Exports FAI FloorSpace GovtRev IP Lending Property Rail Retail	0.53	0.65	0.33
13	AirPassengers ConsumerIndex Electricity Exports FAI FloorSpace GovtRev Highway IP Lending Property Rail Retail	0.54	0.65	0.34
14	AirPassengers ConsumerIndex Electricity Exports FAI FloorSpace GDP GovtRev Highway IP Lending Property Rail Retail	0.56	0.67	0.38

Note: For each number of indicators from 1 to 14, the combination shown has the lowest weighted-average RMSE. RMSEs in sample and out-of-sample are from regressing import growth on the first principal component of the list of indicators shown. Our preferred index is the 8-indicator set, which minimizes in-sample and out-of-sample fit. (All regressions also include the change in the real exchange rate). See also notes to Table 2.

Our best combination with four indicators then adds industrial production, which improves in-sample performances with only a very modest deterioration in out-of-sample performance and lowers average RMSE to 0.51.

Our five-indicator combination adds lending volumes, which leaves in-sample fit unchanged, but achieves a modest improvement in out-of-sample fit for an overall weighted average RMSE improvement to 0.50.

From that point, our results suggest that as we add indicators, fit no longer improves much and eventually deteriorates. Our six-variable combination adds the property indicator, but results in essentially the same weighted average RMSE. Similarly, our seven and eight variable combinations add fixed asset investment and retail sales respectively, with average RMSE remaining at 0.50. The best-fitting 9-indicator combination adds lending, and receives the same rounded score of 0.50, but its true value is modestly higher than the true value of our best-fitting 8-indicator combination.

Given that we are interested in including as many indicators as possible without sacrificing goodness of fit, we therefore choose the best-fitting 8-indicator combination as our preferred specification. This combination includes the consumer index, electricity, exports, fixed asset investment, floor space, industrial production, rail freight, and retail sales.

Henceforth, we label this preferred index based on these eight indicators as the China Cyclical Activity Tracker (C-CAT). A comparison of the C-CAT with the earlier sequential approach is informative. That approach (Table 3) found the best overall fit with 6 indicators—and it turns out that, according to Table 4, that set of indicators has the best fit of *any* 6-indicator combination. However, the 7 and 8 indicator lists add different variables. The 8-indicator set that maximizes overall fit adds fixed asset investment and retail. Those two indicators did relatively poorly on their own but presumably provided different information.

Note that our overall best set of indicators fails to include GDP. Indeed, GDP is never chosen in a best-fitting combination that does not include all indicators. The all indicators set with GDP included has a weighted-average RMSE of 0.56. This is modestly worse than the 13-variable combination of all indicators except GDP, which a weighted RMSE of 0.54. So while GDP on its own is somewhat informative, it provides no additional information on activity that is not already in the alternative indicators.

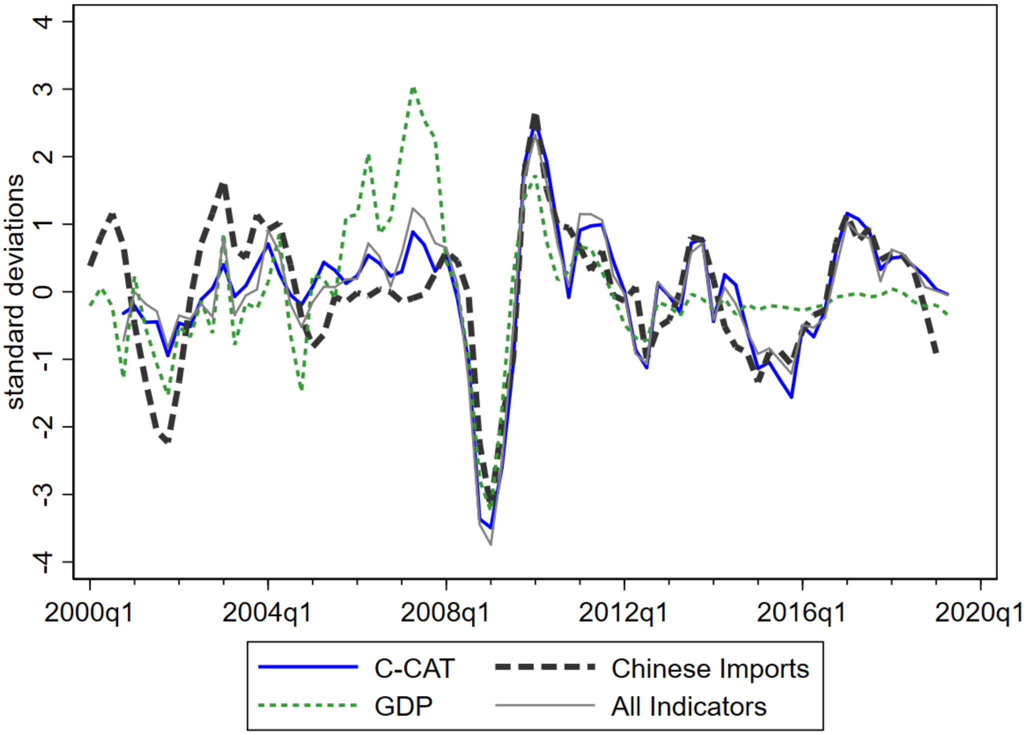
Overall, the comparison of Tables 3 and 4 highlight a number of results: First, while we observe some discrepancies, the qualitative set of individual indicators that perform best is relatively stable. Indicators such as electricity, exports, rail, and industrial production seem to be ones that one would always want to include. Second, the all-indicators combination does well enough that one could rationally chose to use that indicator, letting the data speak solely through the weights chosen in generating the principal component for the activity index. The main advantage of the all-indicators combination is that, while it might not be optimal, it avoids concerns about choosing indicators that just happened to fit well in a particular sample.

Indeed, while we were able to construct combinations that were more parsimonious and outperformed the all indicators activity index, this discrepancy should not be exaggerated. Figure 3 displays the all indicator index, our preferred C-CAT, and reported GDP, as well as exports to China and Hong Kong.

All of the activity indices tend to move closely together, and (by construction) tend to move closely with movements in Chinese imports. Yet, it is apparent that GDP is the most different from the others. Indeed, prior to 2008, GDP shows little comovement with imports. The relationship changes markedly from 2008 through 2013, when GDP and imports comove quite closely. But since 2013, the comovement disappears. Import growth varies about as much as it did prior to the crisis. But GDP growth is extremely flat, with hardly any volatility.

In contrast, China's imports and the C-CAT comove closely throughout the sample, although they do diverge at the end, when imports drop quickly. The C-CAT, like imports, is about as volatile in the period since 2013 as it was prior to the 2008 global financial crisis. Like imports, the C-CAT showed notably slower cyclical growth in 2015—at a time when China was suffering substantial capital outflows and growth was a concern. Imports and the C-CAT turn substantially positive by early 2017, following a loosening in credit and debt. Since then, Chinese authorities began to clamp down on credit and debt, and growth slowed.

Figure 3: **Indicators and Imports**



Note: All series represent four-quarter growth rates in economic activity relative to trend, and are normalized to be mean zero and with a unit standard deviation. China imports are measured as real trading-partner exports to China and Hong Kong. C-CAT is our preferred index (based on the analysis in , and is the first PC of eight indicators listed in the text. "All indicators" is the first principal component of all 14 individual indicators (including GDP). All series are measured 2000Q4-2019Q2 except for imports, which runs through 2019Q1. See text for further details.

Thus, imports and the C-CAT not only move closely together, but their movements match anecdotal evidence regarding the pace of economic activity. We conclude that the C-CAT accurately captures true fluctuations in economic activity. GDP growth, however, appears too smooth to be accurate since 2013.

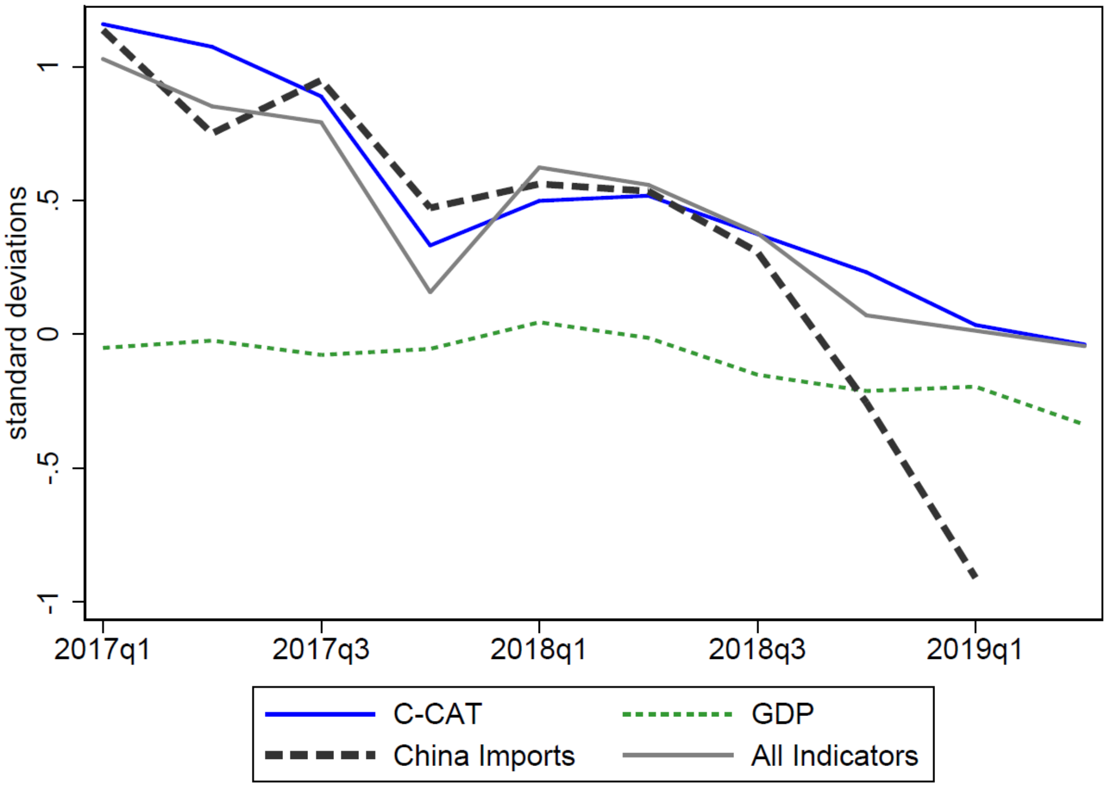
C. Assessing the Recent Period

Since 2013 or so, GDP figures look excessively smooth. Given this, are the most recent figures giving a misleading answer regarding the cyclical slowdown in Chinese economic activity? To evaluate the evidence for this conjecture, we focus on this recent period. Figure 4 plots the same set of indicators for the most recent period since 2017Q1.

Reported Chinese GDP growth has, as noted, been slowing relatively smoothly in recent years. Hence, the readings in Figure 4 are not far from zero. The Chinese imports series, in contrast, has not been smooth. It fell dramatically at the end of 2018—reflecting the trade war (DOTS trade data are only available with a lag, so 2018Q4 is the most recent data available). The CCAT has slowed more noticeably over the past eight quarters (by 1¼ s.d. to a reading of -0.2). Still, it is worth noting that the slowdown was from an above-trend pace, so both GDP and the CCAT are now indicating growth just a little below trend.

Thus, our analysis suggests that, while Chinese cyclical activity has slowed more sharply than suggested by the GDP figures over the past two years, there is no evidence of a collapse in growth to a rate markedly below trend. In that regard, our results also suggest that the GDP figures continue to tell an accurate story about the situation as of the second quarter of 2019.

Figure 4: Experience since 2017



See notes to Figure 3.

All of these series are plotted relative to their own trends. As such, one might conjecture that the higher level of activity suggested by these alternative activity indicators may be driven by declines in trend, due to the recent Chinese slowdown. Were that the case, activity could actually be falling, but not relative to trend, as trend is falling as well.

Although our best estimate is that GDP is currently giving an accurate cyclical signal, this would be cold comfort if it turned out that the decline in trend was even sharper than suggested by the published GDP data.

Our methodology requires detrending at the outset. However, we can compare the movement in the GDP trend to those of our underlying indicators. Trend GDP growth has slowed smoothly and gradually since its post-crisis peak in 2010. The recent year-over-year growth rate of 6.2 percent is 1.1 standard deviations below the full-sample mean of 9.1 percent (see Appendix).

An examination of our individual non-GDP indicators does *not* suggest a sharper slowdown in trend than captured in the published GDP figures. As of 2019Q2, the median estimated trend for our eight individual indicators was -0.7 standard deviations. Some indicators showed an estimated trend decline larger than GDP, including fixed asset investment (-1.3 s.d.) and industrial production (-1.2 s.d). Others showed a smaller decline, including rail freight (0 s.d), new floor space constructed (-0.3 s.d.), and electricity use (-0.6 s.d.).

In sum, Chinese cyclical activity has slowed over the past two years from an above-trend pace to a slightly below-trend pace. Still, there is no evidence of a collapse in growth to a rate markedly below trend—nor do the individual activity indicators suggest a sharper slowdown in trend than the gradual one captured in the published GDP figures.

6. Conclusion

This paper considers imports as a relatively reliable measure of economic activity for countries with poor statistical systems. One virtue is that trade data can be measured using trading partners, so errors should not reflect intentional manipulation. We focus on China, and show how to use exports to China as an externally-verified indicator of economic activity that can be used to assess the reliability of Chinese indicators, including GDP.

With our metric, Chinese statistics, including GDP, became more reliable over time—though the evidence for GDP is more mixed at the end of the sample. Nevertheless, GDP itself adds little information relative to the first principal component of other sets of indicators. Indeed, our preferred set of indicators (the first principal component of the following eight individual indicators: the consumer index, electricity, exports, fixed asset investment, floor space, industrial production, rail freight, and retail sales, which we designate the China Cyclical Activity Tracker) is relatively parsimonious yet provides an accurate assessment of economic activity both in sample and out-of-sample.

We conclude with several caveats. First, imports are an imperfect measure of activity and may underweight certain activities, notably services and other non-tradable sectors. Still, our preferred activity factor includes both relatively narrow indicators (like rail freight) and broader ones (such as air passenger volume and retail sales). Moreover, even if imports or the activity factors are imperfect, there is no reason to think they are necessarily inferior to GDP alone.

Second, even for the pre-2008 period—when GDP is a poor fit of our Chinese economic activity proxy—we cannot say for sure whether GDP was manipulated, or merely limited in its coverage. If manipulation were rampant, we would expect it to be more prevalent during periods of exceptionally high or low economic activity, as data might be changed to more closely meet trend output goals. In that case, measured variation would still reflect true variation, but it would be dampened. In that sense, we cannot say whether the level and variability in GDP are accurate. Rather, we focus on the consistency of the signal over time.

Finally, as China’s economy and statistical system continue to evolve, indicators that do well historically might do less well going forward. Nevertheless, it is reassuring that our core set of indicators performs well across our two sample periods.

References

Bai, J., Perron, P., (1998), “Estimating and testing linear models with multiple structural changes,” Econometrica 66, 47–78.

Bai, J., Perron, P., (2003), “Computation and analysis of multiple structural change models,” Journal of Applied Econometrics 18, 1–22.

Bradsher, Keith (2012). “China Data Mask Depth of Slowdown, Executives Say,” New York Times, June 23. http://www.nytimes.com/2012/06/23/business/global/chinese-data-said-to-be-manipulated-understating-its-slowdown.html?pagewanted=all

Chen, Wei, Xilu Chen, Chang-Tai Hsieh and Zheng Song, (2019), “A Forensic Examination of China’s National Accounts, NBER Working Paper No. 25754, April.

Clark, Hunter, Maxim Pinkovskiy, and Xavier Sala-i-Martin (2018).“China's GDP Growth May be Understated.” China Economic Review.

Dawson, John W., Joseph P. DeJuan, John J. Seater, and E. Frank Stephenson. "Economic Information versus Quality Variation in Cross-Country Data." The Canadian Journal of Economics / Revue Canadienne D'Economique 34, no. 4 (2001): 988-1009. http://www.jstor.org/stable/3131934.

Fernald, John, Israel Malkin, and Mark M. Spiegel, (2013), “On the Reliability of Chinese Output Figures,” FRBSF Economic Letter, 2013-08, March 25.

Fernald, John & Edison, Hali, and Loungani, Prakash, (1999), "Was China the first domino? Assessing links between China and other Asian economies," Journal of International Money and Finance, Elsevier, vol. 18(4), pages 515-535, August.

Hansen, Peter Reinhard and Allan Timmermann, (2012), “Choice of Sample Split in Out-Of -Sample Forecast Evaluation,” EUI Working Paper NO. ECO 2012/10.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil, (2012), “Measuring Economic Growth from Outer Space,” American Economic Review, 102(2), 994–1028.

Holz, Carsten A. (2008). “China’s 2004 Economic Census and 2006 Benchmark Revision of GDP Statistics: More Questions than Answers?” The China Quarterly, March.

Holz, Carsten A. (2013). “Chinese Statistics: Output Data"

Holz, Carsten A. (2003). “‘Fast, Clear and Accurate:’ How Reliable Are Chinese Output and Economic Growth Statistics?” The China Quarterly, no. 173 (March 2003): 122-63.

Nakamura, Emi, Jon Steinsson, and Miao Liu, (2014), “Are Chinese Growth and Inflation Too Smooth?: Evidence from Engel Curves,” mimeo, Columbia University, January.

Noble, Josh (2015). "Doubts rise over China's official GDP growth rate." *Financial Times*, September 16, 2015. http://www.ft.com/intl/cms/s/0/723a8d8e-5c53-11e5-9846-de406ccb37f2.html#axzz3lvULIfcE (accessed September 21, 2015).

Owyang, Michael T. and Hannah G. Shell (2017). "China's Economic Data: An Accurate Reflection, or Just Smoke and Mirrors?" *The Regional Economist*, Federal Reserve Bank of St. Louis, Second Quarter, pp. 7-12.

Pesaran, Hashem, Andreas Pick and Mikhail Pranovich (2013) "Optimal forecasts in the presence of structural breaks" Journal of Econometrics, 177(2) 134-152

Pinovsky, Maxim and Xavier Sala-i-Martin (2016). "Lights, Camera, … Income! Illuminating the National Accounts-Household Surveys Debate." *The Quarterly Journal of Economics*, 131 (2): 579-631, May.

Sharma, Ruchir (2013). "China's Illusory Growth Numbers." Wall Street Journal, October 31, 2013.

Sinclair, Tara (2012). "Characteristics and Implications of Chinese Macroeconomic Data Revisions." Manuscript, George Washington University. http://www.gwu.edu/~iiep/assets/docs/papers/Sinclair_IIEPWP2012-09.pdf

Stock, James H., and Mark W. Watson. 2016. "Dynamic Factor Models, Factor Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." In Handbook of Macroeconomics, Volume 2A, edited by John B. Taylor and Harald Uhlig. Amsterdam: North-Holland.

Subramanian, Arvind, (2019), "India's GDP Mis-estimation: Likelihood, Magnitudes, Mechanisms, and Implications," CID Faculty Working Paper No. 354, June.

Wikileaks (2007). http://wikileaks.org/cable/2007/03/07BEIJING1760.html.

Wright, Jonathan H. (2017). "Optimal Seasonal Filtering." Manuscript, Johns Hopkins University.

Wu, Harry (2014). "China's Growth and Productivity Performance Debate Revisited - Accounting for China's Sources of Growth with a New Data Set." Conference Board Working Paper.

Appendix:

1. Data sources

The chart below shows the raw data we used in the paper. All data were accessed in April 2019, mainly from CEIC Asia database.

Series	Description	Source
Electricity	Electricity production, Billions of kilowatt hours	National Bureau of Statistics (CEIC series 3662501)
Rail	Railway freight traffic, millions of tons	China Railway Corporation, National Railway Administration (CEIC series 12915101)
Lending	Bank loans, billions of RMB	The People's Bank of China (CEIC series 7029101)
Property	Real estate investment (Residential bldgs.), millions of RMB	National Bureau of Statistics (CEIC series 3948701)
Air passengers	Air passenger traffic, millions of persons	Civil Aviation Administration of China (CEIC series 12916401)
Exports	Exports (FOB basis), millions of US dollars	General Administration of Customs (CEIC series 5823501)
Consumer Index	Consumer Expectation Index	National Bureau of Statistics (CEIC series 5198601)
Floor space	Floor space started, thousands of square meters	National Bureau of Statistics (CEIC series 3963901)
Raw materials	Index of raw materials supply, derived from a survey of managers from 5000 companies. Respondents are asked for views on adequacy of supplies of raw materials.	The People's Bank of China (CEIC series 8003501)
Retail	Retail sales of consumer goods, billions of RMB	National Bureau of Statistics (CEIC series 5190001)
Industrial Production	Value added of industry, YoY %	National Bureau of Statistics (CEIC series 3640701)
Highway	Freight carried, highway, Ton mn	China Economic Monitoring & Analysis Center, NBS (CEIC series 12915201)
Government revenue	Billions of RMB	Ministry of Finance (CEIC series 4331701)
FAI	Fixed asset investment, billions of RMB	National Bureau of Statistics (CEIC series 7872901)
GDP	Real GDP index, available as 4-quarter growth rates	National Bureau of Statistics (CEIC series 1692001)

Exchange rate between RMB and USD		Bloomberg
China imports from Hong Kong and China imports from World by Harmonized System category	Quarterly sums in USD mil	General Administration of Customs (accessed through CEIC)
Hong Kong imports from China and Hong Kong imports from World by SITC category	Quarterly sums in HKD mil	Census and Statistics Department (accessed through CEIC)
World exports to Greater China		IMF Direction of Trade Statistics (accessed through CEIC)
US export price indices		BLS (accessed through Haver)

2. Calculation of export price index

We deflate nominal exports to Greater China, defined as world exports to China and Hong Kong minus exports between China and Hong Kong, using an export price deflator calculated from US export price indices, nominal import data for China and Hong Kong, and nominal world export data to China and Hong Kong. First, we calculate separate export price indices for China and Hong Kong. The export price index for each region is equal to the weighted sum of 4-quarter log changes in US export price indices for a number of export categories, where the weights are equal to the category's share of total nominal imports in that region and time period. This can be expressed as

$$\Delta p_t^i = \sum_{k=1}^{n} w_{k,t}^i \, \Delta p_{k,t}^{USex},$$

where Δp_t^i is the export price index for region i in time t, $w_{k,t}^i$ is category k's share of region i's total nominal imports in time t, and $\Delta p_{k,t}^{USex}$ is the 4-quarter log change in the US export price index for category k in time t.

The final price index for exports to Greater China is the weighted sum of the China and Hong Kong export price indices, weighted by each region's share of world exports to greater China in time t.

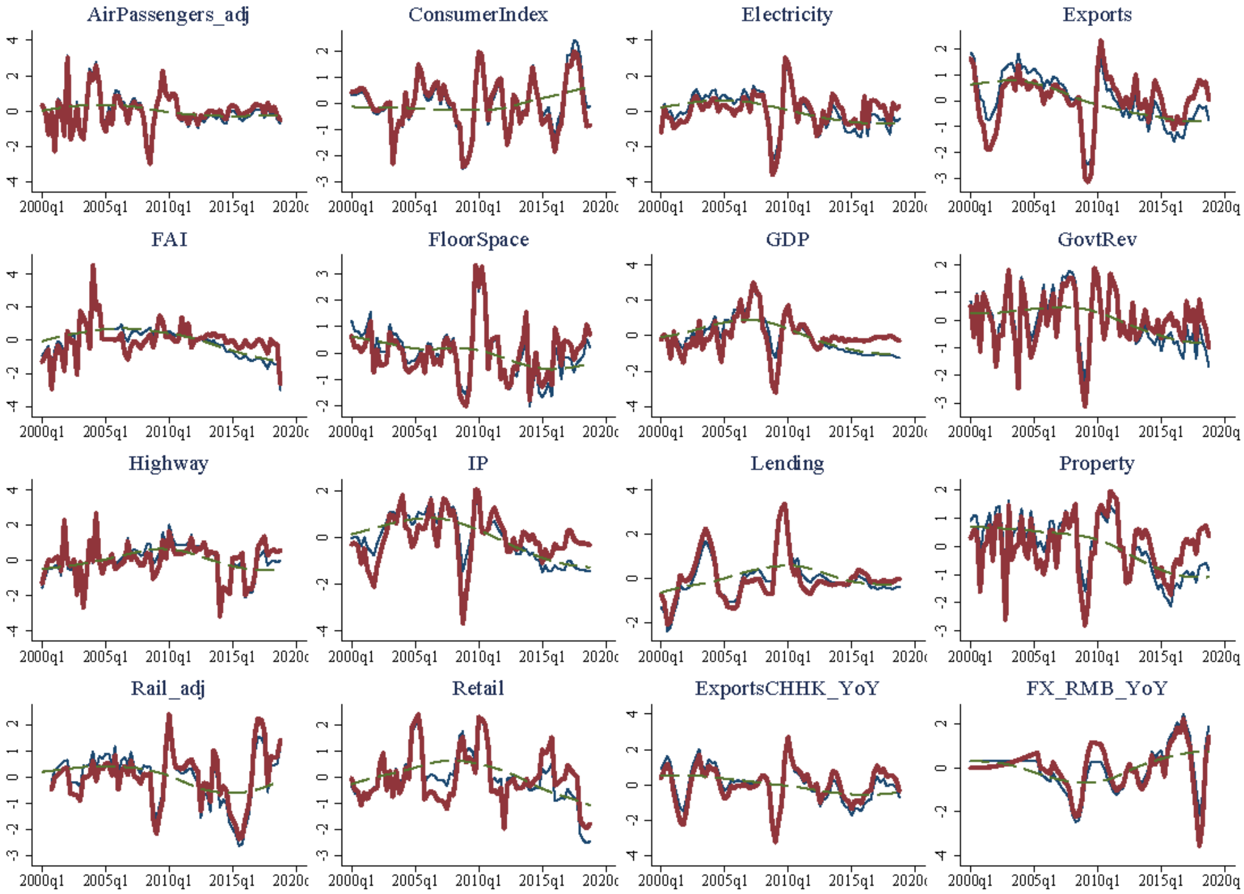
3. Other adjustments

Monthly proxy series converted to quarterly via summing over the quarter.

Missing observations around Chinese New Year:

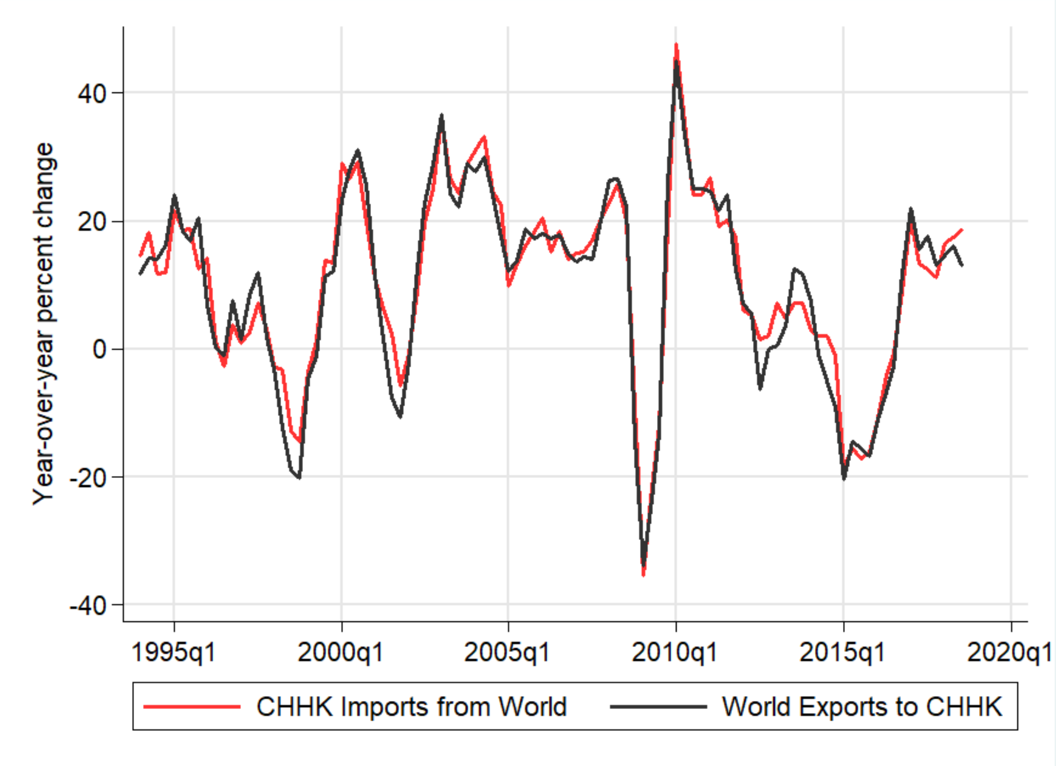
- Electricity missing January and February starting January of 2016. Retail missing January and February starting 2012. Filled in missing January and February values with the March value for years they are missing.
- Rail shipments series has a break in level in January 2005. We adjusted the series by splicing. The splicing adjustment factor comes from regressing log level Rail shipments on the date and a dummy variable for the "post-break" dates in a four-year window around the break (January 2003 to January 2007). The splicing adjustment factor is the reciprocal of the exponentiated coefficient on the dummy variable in this regression and is applied all observations in the Rail series post-break.
- Li: We use the adjusted rail data rather than the raw data.

Figure A1: Individual Indicators



Note: All data are year-over-year percent changes, normalized to be mean zero and unit standard deviation. Vertical axis units are standard deviations. Thin blue line is raw data, normalized; thick red is filtered, normalized; green dashed line is biweight trend (24 quarters). Data run from 2000Q4 to 2019Q2 except for Exports to China and Hong, which end in 2018Q4.

Figure A2: Exports to China Versus China's Imports


Year-over-year percent change
1995q1
2000q1
2005q1
2010q1
2015q1
2020q1
CHHK Imports from World
World Exports to CHHK