Why 1990 international Geary-Khamis dollars cannot be a foundation for reliable long run comparisons of GDP

BY
Liam Brunt AND Antonio Fidalgo

This series consists of papers with limited circulation, intended to stimulate discussion
Why 1990 international Geary-Khamis dollars cannot be a foundation for reliable long run comparisons of GDP

Liam Brunt
Norwegian School of Economics

Antonio Fidalgo
Fresenius University of Applied Sciences

Abstract. Using a large, new dataset of agricultural prices and quantities for many countries and regions, we create five new international Geary-Khamis pounds – for 1870, 1845, 1775, 1705, and a superior chained series. We show that estimated levels and changes in output per worker look very different – more extreme – using 1705 international pounds and 1990 international dollars, compared to all other series; growth rates appear substantially higher using 1990 international dollars. In short, out-of-sample baskets and/or prices create extremely unreliable output estimates. We also show that individual country price indices (rather than international indices) can generate substantially different estimated growth rates.

Keywords: Geary-Khamis prices, economic growth, international comparisons, price indices.
JEL classification: C43, C82, N10, N50.

1. Introduction. Decades of research by dedicated scholars now permit us to make international comparisons of output and productivity going back hundreds of years. A great deal of data for many countries and regions has been unearthed and the methodology for making such comparisons has been firmly established in the context of the International Comparison Program (ICP) and the Penn World Tables (PWT). Unfortunately, the most appropriate methodology has not been implemented with historical data. Therefore, current long run estimates are biased, inconsistent and – frankly – confusing (as we explain in detail in a moment). You can think of there being two problems: a time series problem and a cross section problem. The first (time series) problem is that many comparisons rely on one...
benchmark set of prices to weight the outputs produced in numerous benchmark years, often far removed temporally from the price benchmark. The classic example here is Maddison’s use of the 1990 international dollar to take us back to the time of Christ (Maddison, 2001). We would ideally be using a price benchmark that is closer to our period of interest. Indeed, in a perfect world we would have price benchmarks for every output benchmark and we would chain the prices, as we do in modern national income accounting. The second (cross section) problem is that valid cross-country comparisons require a set of international prices; it is simply not valid to compare GDP levels across countries using their own prices converted at the par of exchange (or nominal exchange rate). Of course, if you cannot make valid comparisons of the relative levels across countries then you cannot make valid comparisons of the relative growth rates because changes in the levels over time imply certain growth rates for each country.

The fundamental stumbling block is that no one has constructed historical Geary-Khamis (GK) international prices – as used in the ICP – to value country outputs, owing to the sheer amount of work required to construct multiple global baskets and match them with appropriate prices. This is why researchers have defaulted to using one price benchmark, such as 1990 international dollars. But this approach does not solve the comparison problem: it merely maximizes the time series problem for the sake of minimizing the cross section problem. Thus it trades one bias (the fact that prices are unchanging and temporally far removed) for another (the fact that national prices do not create a consistent basis for multilateral comparisons). Note that there is no evidence that this is a sensible strategy. It might be better to follow exactly the opposite strategy: minimize the time series problem at the cost of maximizing the cross section problem. That is, use a set of chained prices that is frequently rebased, but impose the same set of prices on all countries. No one has ever examined the “loss function” from implementing these two approaches – which one leads to greater inaccuracy? In fact, it has not been possible to answer this question until now because there is no set of real international prices that we could use as a benchmark (which we need to gauge the degree of inaccuracy of the other approaches).

In this paper, we rectify this lacuna in the literature by the simple expedient of constructing historical international GK prices. That is to say, there is nothing clever in this paper – just an awful lot of hard work. We put together a large, new data set by scouring the secondary literature and spending a lot of time in libraries and archives collecting new data; and we then constructed many different GK price indices by spending a lot of time coding. We are not making any innovative conceptual points because the theory is already well
understood. What has never been shown is just how big the biases are, empirically, when we default to these problematic assumptions instead of collecting and processing data. We show that taking the most theoretically satisfactory approach makes a very big difference to the estimated growth rates.

Unfortunately, the absence of historical GK prices has also made it impossible to compare the output and productivity estimates of different researchers because they are not made on the same basis. Hence we applied the term “confusing” above. For example, someone may compare the levels and growth rates of output per head in Britain and America using nominal exchange rates or PPP exchange rates (Lindert and Williamson, 2016); and someone else may compare the growth rates using own-country price indices (Broadberry, 2003). These may give three different answers and none of them is any more “correct” than any of the others because none of them is theoretically well founded. In fact, we show that these three approaches do give three different answers when estimating output levels and growth rates in the US and UK – and they are not only different to one another, but also different to a comparison using the (more theoretically justifiable) chained GK prices. Although we present our results here in terms of GK prices, we can recast everything in terms of EKS prices and the same arguments – and broadly the same results – pertain.

In section 2, we explain the intricacies of the index number problem in more detail. This is worthwhile because the severity of the problem is really unexpected, unless you have had the misfortune to grapple with index number problems yourself. In section 3, we discuss our data in more detail. In section 4, we present numerous examples from our data of very significant differences in estimated growth rates arising from the use of different indices. In section 5, we conclude.

2. The index number problem. The time series problem is explained very nicely by Coelli et al. (1998). Suppose that you create an output series using terminal prices as weights – which is effectively what you do if you cast everything in 1990 international dollars. Then your output series will appear to grow very fast when goods that have a high terminal price have greatly increased in quantity; and your output series will appear to grow very slowly when goods that have a low terminal price have greatly increased in quantity. If relative prices were the same at the beginning and the end of the period then this is not a problem – there will be no bias. But if the high terminal prices have also grown a lot relative to other prices (for example, there was a positive demand shock for food) then your output index will be overestimated. And if your low terminal prices have shrunk a lot relative to other prices
(for example, there was a positive supply shock for industrial goods) then your output index will be underestimated. Given that long run estimates of output and growth deal with the process of industrialization – which is precisely why they are of general interest – you would imagine that relative prices have changed a lot. Therefore, it is highly likely that pricing everything in 1990 international dollars (i.e. taking a set of relative prices that is 100 to 300 years removed from our period of study) will generate a lot of bias. To put this in perspective, suppose that we worked Maddison’s analysis in the opposite direction. Collect prices and quantities for the year 0CE from the empires of ancient Rome, the Mayans and the Han dynasty in China; construct a Year 0 international denarius; then use this to value country outputs through to 2018. Do we think that the biases are likely to be large enough to be seriously problematic when making international comparisons of the levels and growth rates of GDP?

Now consider the cross section problem. This actually breaks down into several separate problems – real versus nominal exchange rates, the choice of a price basket, and the choice of a particular price vector. We can convert prices from one country into the prices of another country using the nominal exchange rate. But this is quite arbitrary because the exchange rate is determined simply by the supply and demand for currency, which is just a function of the relative volumes of imports and exports. Valuing all goods – non-traded as well as traded – at the nominal exchange rate makes little sense, particularly since the absolute trade volumes may well be small compared to total output in the two countries. Hence economists construct PPP exchange rates, essentially inferring an exchange rate by valuing a basket of goods in the two countries at the two sets of prices (and then expressing them as a ratio). This is a good guide to how much it actually costs to purchase goods, so really captures the relative cost of living in each country. So far, so good. But what basket should we choose for the prices? An English basket in 1775 would have a lot of wheat, whereas a Chinese basket would have a lot of rice. It is likely that wheat was relatively cheap in England (which is why the English consumed a lot of it) and rice relatively expensive (and vice versa for China). So the basket choice will affect the comparison. Using the English basket would make China look like an expensive place to live; using the Chinese basket would make England look like an expensive place to live. GK prices address this problem by effectively estimating a weighted average price for each good, using country shares in world output to weight the country prices (so the basket is the world consumption basket).

Going back to the time series problem, we can say that every relative price change creates a bias and many of these will be offsetting. Then the big question is: what is the
overall (net) bias? Although economic theory allows us to hypothesize about the direction and magnitude of the bias, it is a purely empirical question as to what the overall effect turns out to be. A key component of this paper is to quantify this bias for the first time. At the same time, we do not want to solve the time series problem by introducing a cross section problem — such as employing an arbitrary method of price conversion. We therefore compute international pounds for 1705, 1775, 1845 and 1870. (It makes little sense to calculate international dollars for periods when the dollar does not exist. Therefore we cast everything in international pounds, since the UK was the leading economy through much of this period. In fact, the international currency is only a numeraire and so it makes no practical difference at all if you call it an international dollar or international pound.) We then chain together these price benchmarks to get a price index for the international pound, so that we can then work in “real international pounds”.

If we graph these various price indices then it will be immediately obvious that our choice of index is definitely not innocuous. Figure 1 below harks back to the comparison of Britain and America that we used in the introduction as our motivational example. Take $100 of nominal output in 1850. (So for example, choose an output basket and select an appropriate volume for this basket so that it would be valued at 100 1990 international dollars.) Suppose that this was the value of output produced by one worker, for example. You would like to know how much labor productivity growth has occurred since 1700. You know the nominal value of the output of a worker in 1700 (you have prices for 1700). But in order to compare labor productivity in 1700 and 1850, you need to value the 1850 output in the same prices as the 1700 output (i.e. you need to deflate the 1850 benchmark). Figure 1 shows what happens if you employ different price indices to make this calculation. If you deflate the 1850 nominal output by 1990 international dollars then you get the top line: since 1990 international dollars are unchanging, the real value of the 1850 output cannot change. But if you instead use the US price index then you get the bottom line: the steep decline in the purchasing power of the dollar during and after the American Revolution implies that the real value of a $100 basket in 1850 was much less in 1700. In fact, using the US dollar is problematic because it did not exist until 1785. Since American colonial currencies were based on Sterling before the Revolution, an obvious strategy would be to take the nominal value of output in 1700 and then use the English price index to convert this to an 1850 value; this is the line above the 45-degree line. Finally, you could use chained “international pounds”, which is the most theoretically sound approach and generates the 45-degree line. (The fact that it is 45 degrees is just an accident, given the dimensionality of this particular
graph.) The incredible divergences between these price series generate equally large
divergences in the estimated change in real output per worker (given that we could
reasonably use any or all of these series to infer the change in real output, by using any or all
of them to deflate the change in nominal output). This is the time series problem writ large.

Figure 1. Estimated real US output/worker, using alternative price baskets (1850=100).

The time series problem that we just illustrated arises for any single country (in this
case, America). But now recall that our original remit was to compare two countries (i.e.
Britain and America). If you converted the nominal US incomes into pounds sterling using
the par of exchange, then you would take $4:55 to be equal to £1 in 1791 and $5:59 in 1870
(according to www.measuringworth.com). But if you used PPPs, then you would take $0.84
to be equal to £1 in 1791 and $0.85 in 1858 (Lindert and Williamson, 2016, table D-4). These
would clearly give you very different estimates of the relative levels of output in the two
countries, and they would generate different trends. This is the cross section problem write
large.

Given that there are numerous ways of tackling the time series problem (such as using
domestic or international prices, each constructed in various different ways) and numerous
ways of tackling the cross sectional problem (such as using nominal or PPP or par rates of
exchange), there are many possible ways to make an international comparison of growth in
two countries. And, as figure 1 indicates, the results will be highly dependent on the methodological choices that the researcher makes. So we should not be surprised if two researchers reach very different – even contradictory – answers using the same data. They may both be “right”, given the methodologies that they have adopted.

We know from theory and modern applied work that constructing a real numeraire currency – such the real international pound – is the preferred solution to the multilateral comparison problem. Now if we base our real output estimates on real international pounds – rather than on the 1990 international dollar or on real local currencies – then how much difference does it make? We quantify this directly by estimating agricultural output per worker for 70 countries and regions at the four benchmark dates (1705, 1775, 1845, 1870), and tracking changes over time, using all our national and international price series, as well as using 1990 international dollars. We present these estimates in section 4 below. But, before that, we describe our data set in section 3.

3. The data set. The problem with constructing GK prices is that you need a complete set of output quantities, in addition to the complete set of prices, because the output quantities are used to weight the prices. It is precisely because the data demands are so high that no one has ventured upon this exercise before. We therefore had to build the data set from the ground up, collecting data on the volume of every output (46 products in total) in every country or region of our analysis (around 70 in total). It is not feasible to tackle the whole economy – the data are simply unavailable for most sectors in most countries – so we restricted our analysis to the agriculture only. In fact, we quantified only the arable sector within agriculture because there are essentially no data on animal outputs before 1870 (and only scattered data thereafter, in fact). We believe that it is better to examine a more restricted set of outputs using hard evidence than to broaden the analysis but have to create most of the data ourselves (through assumptions or extrapolations or questionable inferences). Since arable agriculture constituted by far the biggest sector in most economies through most of our period, we believe that our restricted focus still facilitates a reasonable test of the impact of using alternative price series to estimate real output across countries and over time. Our geographical coverage is very wide, as mapped in figure 2 below.
Since our data requirements are quite demanding, we lack observations for most years. Hence we concentrate on four benchmark years (1701, 1775, 1845 and 1870). Since we are focusing on long run growth comparisons – rather than short run fluctuations – we believe that this is a sound approach.

There were many challenges in terms of definitions. For example, the definition of many regions changed over time – the US being the most obvious example, growing from a cluster of 13 colonies on the eastern seaboard of North America in 1705 to a nation state stretching to the Pacific Ocean by 1870. In fact, one reason that we present results mostly based on real output per worker is that total output was changing radically as countries’ effective land area increased dramatically (as in Australia, Argentina, Canada, China and Russia, to name just a few).

There were also many challenges in terms of conversions, since we were dealing with a lot of archival material that was recorded in local and/or obscure units. In the interests of thoroughness and openness, the sources of our data – and the ways in which they were processed – are fully documented in Brunt and Fidalgo (2018). Readers can also view our data and reproduce any of the graphs in this paper – or others of their own choosing – using our prototype productivity web calculator at https://bruntfidalgo.shinyapps.io/try_shine/.
4. Assessing the impact of using alternative price indices. A first key result of our investigation is that – for most countries – labour productivity estimates (both levels and changes) are very similar when using 1775 international pounds, 1845 international pounds, 1870 international pounds or real (i.e. chained) international pounds (labeled BF in the figures). But the results are radically different using 1990 international dollars (growth is then estimated to be much faster) or 1705 international pounds (when growth is estimated to be much slower). Figure 3 below demonstrates this for England, our baseline case, whilst figures 4 and 5 use Canada and France to show that this is a typical result. This should really be no surprise: if you choose to use a price basket constructed at one end of your period of observation (or well outside it, in the case of the 1990 international dollar) then you would expect any biases to be maximized. The fact that the 1705 international pound generates much lower estimated growth rates, whilst the 1990 international dollar generates much higher estimated growth rates, should also be no surprise: since relative prices were obviously moving substantially over the period, it is natural that baskets taken from either extreme (early or late on the period of analysis) will be biased in opposite directions.

Figure 3. Estimated labour productivity in English agriculture, different price baskets.
We now plot Spain in figure 6 below to show that the exact opposite result can also occur: growth may appear lower (in this case, negative) when using 1990 international dollars. This demonstrates that it is not safe to make historical comparisons using 1990 international dollars, as has often been attempted: the results differ markedly – and unpredictably – from estimates using more relevant historical price benchmarks or indices. It is not even the case that the direction of change, or the relative ranking of countries, remains unaffected by using an extreme benchmark, such as 1990 international dollars.
Our next exercise is to compare country growth rates estimated using international pounds or own-country prices. This reveals our second key result: estimated growth rates are sometimes similar, but sometimes very different, when using domestic versus international prices. The examples of Ireland (which is essentially the same for all price indices) and South Africa (where the international and domestic indices move in opposite directions) demonstrate this in figures 7 and 8 below. This result is easily understandable. If the own-country price structure is similar to the international price structure then estimate growth will be the same whether we use domestic or international prices. But if the own-country price structure is very different to the international price structure then the estimated growth may be very different.
Note that big countries are likely to have a price index that moves more similarly to the international price index because they have a big weight in its construction (and, after all, the international price is just a weighted average). Also, the price index of diversified countries is likely to track the international price index more closely because divergent trends are more likely to offset one another. So if you take a small country, such as the Kingdom of the Two Sicilies, which specializes very heavily in one product (wheat), then it is no surprise that its estimated change in real output looks very different if you use national or international real prices; we can see this in figure 9 below, where we present results for
Sardinia (which was part of the Kingdom of the Two Sicilies at this time).

Figure 9. Sardinian agricultural labour productivity: domestic v. international prices.

5. Conclusions. We know how real output levels and changes ought to be estimated over the long run: that is, using a chained index of historical international prices (constructed using some method such as GK, or perhaps EKS) to deflate nominal output values (both across countries and over time). This is theoretically well founded. It also offers the benefit of removing the enormous confusion and uncertainty that arises from having different researchers use different price indices (such as domestic price indices rather than international price indices, or PPPs rather than nominal exchange rates). It also has the significant benefit that – once the data are collected and the calculations made – we will not need to rework the figures again; this is not the case currently, when all the figures need to be rebased every time someone changes the base year of the international dollar.

It turns out that country experiences typically look very different if you use chained GK prices than if you use 1990 international dollars. We show that growth generally appears slower with chained GK prices – although this is unfortunately not true for every country, thus making historical cross-country comparisons based on 1990 international dollars a poor guide to the true state of affairs.

We also show that it is not safe to compare growth rates across countries using country-specific price indices. This latter issue has recently been addressed by Johnson *et al.* (2013) in the context of the PWT; they conclude that country-specific prices indices may be
less biased when using annual data, but more biased over the long run. We show that over the very long run, country-specific price indices can be very biased. Instead of using either 1990 international, or country-specific price indices, we should extend the nascent set of historical GK prices that we have produced to include more benchmarks and more products.

Bibliography.

01/17 January, **Agnar Sandmo**, “Should the marginal tax rate be negative? Ragnar Frisch on the socially optimal amount of work”

02/17 February, **Luca Picariello**, “Organizational Design with Portable Skills”

03/17 March, **Kurt R. Brekke**, Tor Helge Holmås, Karin Monstad og Odd Rune Straume, “Competition and physician behaviour: Does the competitive environment affect the propensity to issue sickness certificates?”

04/17 March, **Mathias Ekström**, “Seasonal Social Preferences”.

05/17 April, Orazio Attanasio, Agnes Kovacs, and **Krisztina Molnar**: “Euler Equations, Subjective Expectations and Income Shocks”

06/17 April, **Alexander W. Cappelen**, Karl Ove Moene, Siv-Elisabeth Skjelbred, and **Bertil Tungodden**, “The merit primacy effect”

07/17 May, **Jan Tore Klovland**, “Navigating through torpedo attacks and enemy raiders: Merchant shipping and freight rates during World War I”

08/17 May, **Alexander W. Cappelen**, Gary Charness, **Mathias Ekström**, Uri Gneezy, and **Bertil Tungodden**: “Exercise Improves Academic Performance”

09/17 June, **Astrid Kunze**, “The gender wage gap in developed countries”

10/17 June, **Kristina M. Bott**, **Alexander W. Cappelen**, **Erik Ø. Sørensen** and **Bertil Tungodden**, “You’ve got mail: A randomized field experiment on tax evasion”

11/17 August, Marco Pagano and **Luca Picariello**, “Talent Discovery, Layoff Risk and Unemployment Insurance”

12/17 August, Ingrid Kristine Folgerø, **Torfinn Harding** and Benjamin S. Westby, “Going fast or going green? Evidence from environmental speed limits in Norway”

13/17 August, **Chang-Koo Chi**, Pauli Murto, and Juuso Välimäki, “All-pay auctions with affiliated values”
14/17 August, Helge Sandvig Thorsen, “The effect of school consolidation on student achievement”.

15/17 September, Arild Sæther, “Samuel Pufendorf and Ludvig Holberg on Political Economy”.

16/17 September, Chang-Koo Chi, Pauli Murto, and Juuso Välimäki, “War of attrition with affiliated values”.

17/17 September, Aline Bütkofer and Giovanni Peri, “The Effects of Cognitive and Noncognitive Skills on Migration Decisions”

18/17 October, Øivind Schøyen, “What limits the powerful in imposing the morality of their authority?”

19/17 October, Charlotte Ringdal and Ingrid Hoem Sjursen, “Household bargaining and spending on children: Experimental evidence from Tanzania”

20/17 December, Fred Schroyen and Karl Ove Aarbu, “Attitudes towards large income risk in welfare states: an international comparison”

21/17 December, Alexander W. Cappelen, Ranveig Falch, and Bertil Tungodden, “The Boy Crisis: Experimental Evidence on the Acceptance of Males Falling Behind”

2018

01/18 January, Øystein Foros, Mai Nguyen-Ones, and Frode Steen, “Evidence on consumer behavior and firm performance in gasoline retailing”

02/18 January, Agnar Sandmo, “A fundamental externality in the Labour Market? Ragnar Frisch on the socially optimal amount of work”

03/18 February, Pierre Dubois and Morten Sæthre, “On the Effect of Parallel Trade on Manufacturers’ and Retailers’ Profits in the Pharmaceutical Sector”

04/18 March, Aline Bütkofer, Julie Riise, and Meghan Skira, “The Impact of Paid Maternity Leave on Maternal Health”

05/18 March, Kjetil Bjorvatn and Bertil Tungodden, “Empowering the disabled through savings groups: Experimental evidence from Uganda”
April, Mai Nguyen-Ones and Frode Steen, “Measuring Market Power in Gasoline Retailing: A Market- or Station Phenomenon?”

April, Chang Koo Chi and Trond Olsen, “Relational incentive contracts and performance”

April, Björn Bartling, Alexander W. Cappelen, Mathias Ekström, Erik Ø. Sørensen, and Bertil Tungodden, «Fairness in Winner-Take-All Markets»

April, Aline Bütkofer, Sissel Jensen, and Kjell G. Salvanes, «The Role of Parenthood on the Gender Gap among Top Earners»

May, Mathias Ekström, “The (un)compromise effect”

May, Yilong Xu, Xiaogeng Xu, and Steven Tucker, «Ambiguity Attitudes in the Loss Domain: Decisions for Self versus Others»

June, Øivind A. Nilsen, Per Marius Pettersen, and Joakim Bratlie, “Time-Dependency in producers’ price adjustments: Evidence from micro panel data”

June, Øivind A. Nilsen, Arvid Raknerud, and Diana-Cristina Iancu, “Public R&D support and firms’ performance. A Panel Data Study”

June, Manudeep Bhuller, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad: «Incarceration, Recidivism, and Employment”

August, Manudeep Bhuller, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad: «Incarceration Spillovers in Criminal and Family Networks”


August, Chang Koo Chi, “An analysis of the two-bidder all-pay auction with common values”

August, Alexander W. Cappelen, Cornelius Cappelen, and Bertil Tungodden, “Second-best fairness under limited information: The trade-off between false positives and false negatives”

20/18  September, Juan Pablo Atal, José Ignacio Cuesta, and Morten Sæthre, “Quality regulation and competition: Evidence from Pharmaceutical Markets”

21/18  October, Orazio Attanasio, Agnes Kovacs, and Krisztina Molnar, “Euler Equations, Subjective Expectations and Income Shocks”

22/18  October, Antonio Mele, Krisztina Molnár, and Sergio Santoro, “On the perils of stabilizing prices when agents are learning”


24/18  November, Ingrid Hoem Sjursen, “Accountability and taxation: Experimental evidence”

25/18  November, Liam Brunt and Antonio Fidalgo, “Why 1990 international Geary-Khamis dollars cannot be a foundation for reliable long run comparisons of GDP”