Roger, Lionel

**Working Paper**

# A replication of "The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis" (Oxford Bulletin of Statistics and Economics, 2014)

Economics Discussion Papers, No. 2019-27

**Provided in Cooperation with:**
Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

*Suggested Citation:* Roger, Lionel (2019) : A replication of "The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis" (Oxford Bulletin of Statistics and Economics, 2014), Economics Discussion Papers, No. 2019-27, Kiel Institute for the World Economy (IfW), Kiel

This Version is available at:
https://hdl.handle.net/10419/194879

Economics
The Open-Access, Open-Assessment E-Journal

**Discussion Paper**
No. 2019-27 | April 03, 2019 | http://www.economics-ejournal.org/economics/discussionpapers/2019-27

# A replication of 'The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis' (Oxford Bulletin of Statistics and Economics, 2014)

*Lionel Roger*

**Abstract**
Macroeconomic data have been shown to vary substantially between sources, especially so for low-income countries. While the impact of data revisions on inference is well documented for cross-country studies, there is no systematic analysis of the robustness of results obtained from time series analysis. This is despite the fact that time series analysis is an integral part of the econometric toolkit of government analysts, and informs policy decisions in many areas of macroeconomics. This study fills this gap for the notoriously controversial aid-effectiveness debate using the statistical framework by Juselius et al. (2014, *Oxf Bull Econ Stat*): by adopting alternative sources of GDP data in 36 sub-Saharan African countries The author finds that results remain robust across datasets in two thirds of the countries, but sometimes drastically change in others. These findings suggest that robustness checks such as those carried out here should become standard procedure for macroeconomic analysis using single-country time series.

(Replication Study)

**JEL** C32 F35 O11
**Keywords** Time-series models; economic growth; economic data; foreign aid

**Authors**
*Lionel Roger*, School of Economics, University of Nottingham, UK,
lionel.roger@nottingham.ac.uk

*The author is grateful to Oliver Morrissey and Markus Eberhardt for their guidance throughout the research process. He also thanks Katarina Juselius, Andreas Fuchs, Heino Bohn Nielsen, Kevin Lee, Christopher Adam, Basile Boulay, Martina Magli, and the participants of the CSAE Conference 2016 at St Catherine's College, Oxford, for their suggestions and comments. Furthermore he thanks Niels Framroze Møller for providing him some of the data and programs underlying Juselius, Møller and Tarp (2014). This work is part of the author´s PhD research supported by the Economic and Social Research Council [award number 1511835].*

**Citation** Lionel Roger (2019). A replication of 'The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis' (Oxford Bulletin of Statistics and Economics, 2014). Economics Discussion Papers, No 2019-27, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2019-27

**Received** December 12, 2019 **Accepted as Economics Discussion Paper** April 2, 2019
**Published** April 3, 2019

# 1 Introduction

In recent years, there have been growing concerns among scholars and policy-makers about the reliability of macroeconomic data, especially for sub-Saharan African countries (see Jerven, 2013c). Partly as a consequence of the low quality of the National Accounts data reported by national statistical offices, there can be large differences between the numbers reported by different datasets such as the Penn World Table (PWT) or the World Development Indicators (WDI): First, data providers may base their series on different vintages of sometimes substantially revised National Accounts data, and second, they often account for shortcomings and gaps in the original data using different methodologies. A growing body of literature is concerned with the impact of such differences on macroeconomic inference (e.g., Ciccone and Jarociński, 2010; Ponomareva and Katayama, 2010; Johnson et al., 2013; Breton, 2015; Breton and García, 2016). Two core findings of this literature motivate the study at hand: First, the GDP series tend to diverge significantly more for countries with low incomes. Second, the results of studies that exploit annual variation in the data seem to be more sensitive to such discrepancies (Johnson et al., 2013). While this finding refers to a limited set of cross-country studies (Ramey and Ramey, 1995; Jones and Olken, 2005; Hausmann et al., 2005), it suggests that time series analyses, which rely entirely on the temporal variation in the data, may be particularly affected by this fragility. On the other hand, they rely entirely on the variation *within* countries, and therefore remain largely unaffected by measurement issues related to purchasing power *across* countries. This may in turn increase their robustness compared to cross-country studies. To my knowledge, there has been no systematic investigation of this issue in the previous literature. The aim of this study is to fill this gap for the aid-effectiveness literature, which increasingly relies on country-specific time series analysis (e.g., Juselius et al., 2014; Gebregziabher, 2014; Juselius et al., 2017; Addison and Baliamoune-Lutz, 2017).

Before assessing the impact on the inference, this study examines the nature and extent of the discrepancies between the employed series, namely the Penn World Table versions 6.3, 7.1 and 8.0[1], and the World Development Indicators (PWT6, PWT7, PWT8 and WDI in what follows). Special attention is given to the growth rates of GDP and its components, which, from a theoretical perspective, are likely to matter most in the context of a time series analysis of economic growth. In the 36 sub-Saharan African countries included in the analysis, there is indeed striking disagreement between these sources. For instance, none of the correlations between the investment growth series in WDI and those in different vintages of the PWT are higher than 0.07. Even between different releases of the PWT, the series tend to change substantially: for instance, the GDP growth rates reported in PWT6 and PWT7 disagree by 3.3 percentage points on average in every year, which is almost as large as the average reported growth rate (3.7% in PWT6). A decomposition of the divergence between these two series reveals that the largest part of it is explained by changes in the underlying price estimates of the GDP expenditure shares.

For their far-reaching policy implications and dependence on often unreliable data, the results of the literature on foreign aid effectiveness arguably deserve special scrutiny. The conclusions of this literature – in the past mainly obtained from cross-country studies – have previously proven particularly susceptible to even modest changes in the dataset (Easterly et al., 2004; Roodman, 2007) or seemingly subtle nuances in the construction of

[1]The most recent version of the PWT at the time the analysis was carried out.

variables (Van de Sijpe, 2013). In recent years however, an increasing number of studies made use of multiple equation time series models in order to identify the dynamics of foreign aid within countries over time (e.g., Osei et al., 2005; Gebregziabher, 2014; Juselius et al., 2014, 2017; Addison and Baliamoune-Lutz, 2017; Bwire et al., 2017; Mascagni and Timmis, 2017). There are substantial merits to this approach: These mostly country-specific studies do not rely on the restrictive assumption of parameter homogeneity (see Eberhardt and Teal, 2011), allowing for each country to have their own dynamics and long-run equilibrium. Furthermore, they can exploit the full potential of the time series properties of the data, such as identifying long-run and short-run dynamics, while being robust to their pitfalls, like spurious regression due to non-stationary variables.

At the same time, however, these studies put particularly high requirements on the quality of the data, while the level of technicality and the intricate process of model specification can act as a barrier when it comes to performing robustness checks. It is therefore understandably not common practice in this branch of the literature to assess the robustness of one's conclusions to alternative datasets; the aim of this study is to assess the potential importance of this omission. As opposed to the cross-country literature, where robustness checks on the main results of a study often only require the implementation of a single regression model for the entire sample, time series methods generally involve careful modelling of each country under investigation. Indeed, most studies limit their attention to a single country. This substantially complicates any systematic investigation of the robustness of the results in this literature. In order to do so nonetheless, I will take as a point of departure the framework provided by Juselius, Møller and Tarp's (2014, henceforth JMT) recent study on the long-run effects of foreign aid on economic growth. It provides well-specified country-specific Cointegrated VAR (CVAR) models for as many as 36 sub-Saharan African countries, all derived within the same, clearly defined econometric framework. I will then take their models to the data from PWT6 (thus replicating JMT's study with the same data they used), PWT7, PWT8 and WDI. As it is at the heart of the philosophy of the CVAR framework (and the time series literature more generally) to '[allow] the data to speak freely' (Hoover et al., 2008, title), I will also explore the impact of changes in the data on the modelling process itself, and how this is reflected in the results.

The empirical analysis then comprises two parts: First, I apply the 36 country-specific models exactly as specified by JMT to the new datasets. If the data were consistent across datasets, one would expect the models fitted around the original data to also be accurate models of the new data, and the results to be the same or at least similar. If the results change however, this can indicate that either the new data tell a different story altogether, or that the model that accurately captures the old data does not apply to the new data (it is misspecified). In the latter case, it may still be the case that the data effectively tell the same story, but that modifications in the statistical model are needed in order to accurately describe the new data. To address this, in the second part, I re-specify the models for each dataset individually using the statistical criteria employed by JMT for the 12 countries where data are available from all datasets considered.

In the first exercise, using JMT's models, for approximately one third of the countries in the sample the results change qualitatively, whereas two thirds remain stable. The second exercise, re-specifying the models, induces somewhat greater divergence in the results, but significantly less so for countries where the results remained stable in the first exercise. Interestingly, the stability of the results for any given country does not appear to be systematically linked to the divergence of the time series across datasets: The datasets

disagree just as much over the levels and growth rates of GDP and the relative movement of its components in countries where the results remained stable throughout the robustness checks, as in those where the results turned out to be particularly unstable. Overall, my results confirm that the differences between standard datasets can often have very substantial effects on the conclusions drawn from time series analysis, and I suggest that robustness checks with respect to the data should be standard practice in this literature.

The remainder of this paper is organised as follows: Section 2 provides a brief overview of the literature. Section 3 introduces the datasets and discusses the extent, nature and origins of their divergence. Special attention will be given to aspects that are particularly relevant in the context of time series analysis. Section 4 summarises the methodology employed by JMT, and in section 5 I specify my replication procedures and present the results of both exercises. Section 6 concludes.

# 2 Related literature

In his 2013 book, 'Poor Numbers', Morten Jerven laments the low quality of National Accounts statistics from developing countries, and warns that this may have serious consequences for researchers' conclusions and policy-makers' decisions (Jerven, 2013c). He argues that, due to a lack of statistical capacity, political will, and changing academic currents, GDP measures quantifying the economic performance of African economies tend to be severely flawed, and often incomplete (see also Jerven, 2011, 2013b,a, 2016). As data providers try and fill in gaps in the data, to restore consistency within series and to establish comparability across countries, large differences between datasets can accrue. A growing literature is concerned with quantifying and explaining these discrepancies, and how they affect the macroeconomic inference that is based on these datasets.

Ram and Ural (2013) show that, between the Penn World Table (version 7.1) and the World Development Indicators 2012, the estimated GDP per capita in 2005 diverges by more than 25% in as many as 33 countries. The differences range from -54% to +66% of GDP in WDI as compared to PWT. The largest relative differences (relative to GDP level) almost exclusively occur in low-income countries, and no developed country exhibits differences of more than 25% of GDP. These differences do not appear to be following an obvious pattern across datasets, in the sense that neither WDI nor PWT systematically reports higher or lower incomes.

Perhaps more striking than the divergence *between* datasets is the divergence *within* different vintages of the same series, a finding that has often been confirmed for the Penn World Tables. Analysing the differences between four different versions of PWT, ranging from version 5.0 (1991) to 6.1 (2002), Ponomareva and Katayama (2010) find substantial divergence in the GDP growth rates. Using each of these datasets, they replicate the influential contribution by Ramey and Ramey (1995) on the link between business cycle volatility and economic growth. The results vary strongly from one dataset to another and the main result – that countries with higher volatility have lower growth – collapses using some versions of the PWT. In a similar exercise, Atherton et al. (2011) replicate Hanushek and Kimko (2000)'s analysis of the link between labour force quality and economic growth using PWT 6.1 and 6.2, also inducing significant differences in the results. Equally using PWT 6.1 and 6.2, Ciccone and Jarociński (2010) assess the robustness of Sala-i-Martin et al. (2004)'s 'agnostic' approach to growth empirics to differences in the data. Again, the differences between the PWT vintages have a strong impact on the results.

In a similar but more exhaustive exercise, Johnson et al. (2013) compare the growth

rates reported in versions 6.1 and 6.2 of the PWT, and replicate 13 major empirical contributions to the growth literature using these datasets. Their analysis shows that the datasets tend to disagree more for countries with smaller size (as measured by GDP) and older benchmark years. In line with what is suggested by the studies discussed above, Johnson et al. (2013) find that some standard results in the growth literature are not robust across different versions of the PWT. But their results also suggest that this fragility is systematically linked to methodological properties of the studies. Crucially, studies relying on relatively long averages (over periods longer than 5 years) are comparatively stable to changes in the underlying data, while analyses exploiting annual variation in the data yield the least robust results. This situation is exacerbated when the analysis refers to poorer (non-OECD) countries. This is of particular relevance for the study at hand, as JMT's analysis, and therefore I, exploit the dynamics of annual data, and focus on some of the poorest countries in the world. Looking at more recent releases of the PWT, the inconsistency does not appear to have decreased, and Breton (2012) finds even larger differences than those observed by Johnson et al. (2013) between versions 6.3 (2009) and 7.0 (2011).

Any comparison between datasets is limited by the fact that there is no straightforward way of assessing the accuracy of one dataset against another. One way of addressing this question is to examine the assumptions and methodologies underlying the datasets, and to assess their validity. For instance, Breton (2015) and Breton and García (2016), show that the main difference between ICP 1996 and ICP 2005, reflected in the Penn World Tables' PPP estimates, are the estimated prices of investment in developing countries. They show that this is mainly driven by changes in the methodology used to estimate prices in the construction sector: ICP 2005 bases its estimates on the prices of inputs, while earlier versions of the ICP estimated the prices of construction projects as such – a more demanding, but arguably more precise procedure. Breton and García (2016) come to conclude that earlier versions of the PWT (versions before PWT7) were superior in this respect. The most recent round of the ICP, based on prices collected in 2011, again differs substantially from the 2005 vintage. Both Deaton and Tten (2017) and Inklaar and Prasada Rao (2017) show that this discrepancy is likely explained by methodological shortcomings of ICP 2005. These led to overstated prices in lower income countries, and consequently overstated differences in incomes across countries.

Pinkovskiy and Sala-i-Martin (2016) take a more data-driven approach at assessing the relative accuracy of datasets. Building on a recent literature that uses human nighttime light emissions as a proxy for economic growth (most notably, Henderson et al., 2012) they provide estimates of the relative accuracy of different vintages of PWT as well as WDI. In contrast to Breton and García (2016), their findings suggest that the 2005 ICP round provides an improvement over earlier rounds of the survey. Nevertheless, their results corroborate the frequent finding that 'newer need not be better', in the sense that PWT 7.1 consistently outperforms the more recent PWT 8.0 and 8.1 series. It further emerges from their analysis that PPP adjustments in general tend to decrease accuracy in terms of growth rates, and they recommend using WDI if the focus is on growth (a point also made by Deaton and Heston, 2010).

In summary, it is a well-established finding that the most commonly used datasets tend to disagree to a substantial degree, and this is more pronounced in poorer countries. There is no conclusive evidence as to which dataset is the most reliable, and a large number of studies have been shown to be fragile across datasets. The time series literature has so far been exempt from such an investigation, a gap that this study aims to fill.

# 3 Data

The datasets compared in this study are three versions of the Penn World Table (6.3, 7.1 and 8.0), as well as the World Bank's World Development Indicators 2015. From each, I use data on GDP and its components, namely Investment, Household Consumption and Government Consumption. All aid data are ODA net disbursements as reported by the Development Assistance Committee (DAC) and published by the OECD, which is the same source and measure JMT use.[2] This section will briefly discuss the key methodological differences and similarities between the GDP datasets, and then explore how this is reflected in the data.

Before discussing the individual GDP datasets, a general point concerning the choice of series needs to be clarified. As mentioned in the previous section, there are arguments against using PPP adjusted series where the main interest is in temporal variation (growth rates). Nevertheless, for various reasons (both conceptual and related to data availability), much of the empirical growth literature does rely on PPP adjusted datasets, most notably the PWT. Our baseline study, JMT, is no exception, and we choose our alternative series such that they are conceptually close to the series they originally employed and have sufficient coverage. Three of the four series we employ (PWT6, PWT7, PWT8) are then PPP adjusted, and one is not (WDI; see next section for discussion). Section 3.3 will provide some insights as to the influence of price estimates on growth rates, which is the main concern with PPP adjusted series in the context of growth studies.

## 3.1 Conceptual comparison of the datasets

Penn World Table 6.3 (Heston et al., 2009) is the dataset underlying JMT's original study, used here as the benchmark dataset for the purpose of replication. The series JMT employ is real[3] GDP at constant 2005 prices, computed using a Laspeyres index (labelled RGDPL). The prices underlying the PPP adjustments in PWT6 are based on ICP 1996 estimates.

Penn World Table 7.1 (Heston et al., 2012) provides the same variable, RGDPL, but conceptually differs from PWT6 in two ways: First, PWT7 exclusively relies on the prices from the newer ICP round 2005. As noted by Breton (2012), it discards all older price data, inducing major differences in the reported growth rates. Second, the underlying concept of consumption is different from that in earlier (and later) versions of the PWT. Instead of differentiating between household consumption expenditure (HCE) and government expenditure, PWT7 uses the concept of *actual individual consumption* (AIC) and *collective government consumption* (CGC). The difference between the two lies in the treatment of goods and services that are consumed by individuals but often paid for by the government, such as healthcare and education. AIC includes these expenditures, whereas HCE only includes such expenditures that are actually being paid for by the individual. While AIC arguably establishes better comparability across different economic systems, it has been discarded in later versions of the PWT as the required data are not readily available for most countries, increasing guesswork (Feenstra et al., 2013a).

The Penn World Table 8.0 (Feenstra et al., 2013b) introduces a wide range of methodological changes. Many of them specifically aim at increasing the consistency across

[2]The aid data can sometimes be subject to revisions, which proved, however, not to have a substantial impact on the results as I will show in section 5.

[3]Note that *real* stands for *PPP adjusted* in the PWT, instead of *in constant prices*.

(future) versions and reducing the amount of speculation in the reported data. As a consequence, 22 countries with particularly poor data coverage have been removed from the database, 3 of which were included in JMT's study, reducing our sample from 36 to 33 countries for PWT8. The new authors also discarded the RGDPL series JMT base their analysis on. I therefore use the conceptually most similar series included in the dataset, labelled RGDPna, which the authors confirm to be the measure most in line with previous versions of the PWT (Feenstra et al., 2015).[4] The series also corresponds to PPP adjusted GDP at 2005 prices, with the difference that the growth rates applied are GDP growth rates taken directly from the underlying national accounts data.[5] Note that the authors do not explicitly report the expenditure shares employed in the study at hand in the PWT release. They do, however, provide the underlying national accounts data which contains this information.

The World Bank's World Development Indicators (The World Bank, 2015) provide two different series of GDP at constant prices - in 2005 US dollars, and in PPP adjusted 2005 US Dollars. While the latter is conceptually closer to the PWT measures JMT and I employ, it only starts in 1990 for most countries. Note that, since all series are in constant prices, the temporal variation in prices is not directly reflected in the growth rates. Differences do, however, arise through the relative valuation of individual expenditure shares. For instance, a very low price for investment may substantially inflate the real GDP estimate in a period when investment is relatively high compared to the other measures (see section 3.3 below). However, there is no compelling reason to use PPP adjusted series in the present analysis other than data availability and conventions in the literature: By its nature, JMT's study (and therefore my analysis) does not rely on differences in income *levels* across countries, as it does not exploit the cross-sectional dimension. I therefore opt against the PPP adjusted series in WDI, in order to retain an already drastically reduced sample of 13 countries.[6]

## 3.2 Relative divergence of the datasets

Figure 1 plots each of the GDP series for 8 selected countries; anticipating the findings of section 5.2, the four countries at the top of the panel are those where JMT's results remain the most stable in our replication exercise, and the countries at the bottom of the panel are those with the least consistent results. In order to illustrate differences in the reported composition of GDP, figure 2 plots the corresponding investment shares for those same countries.

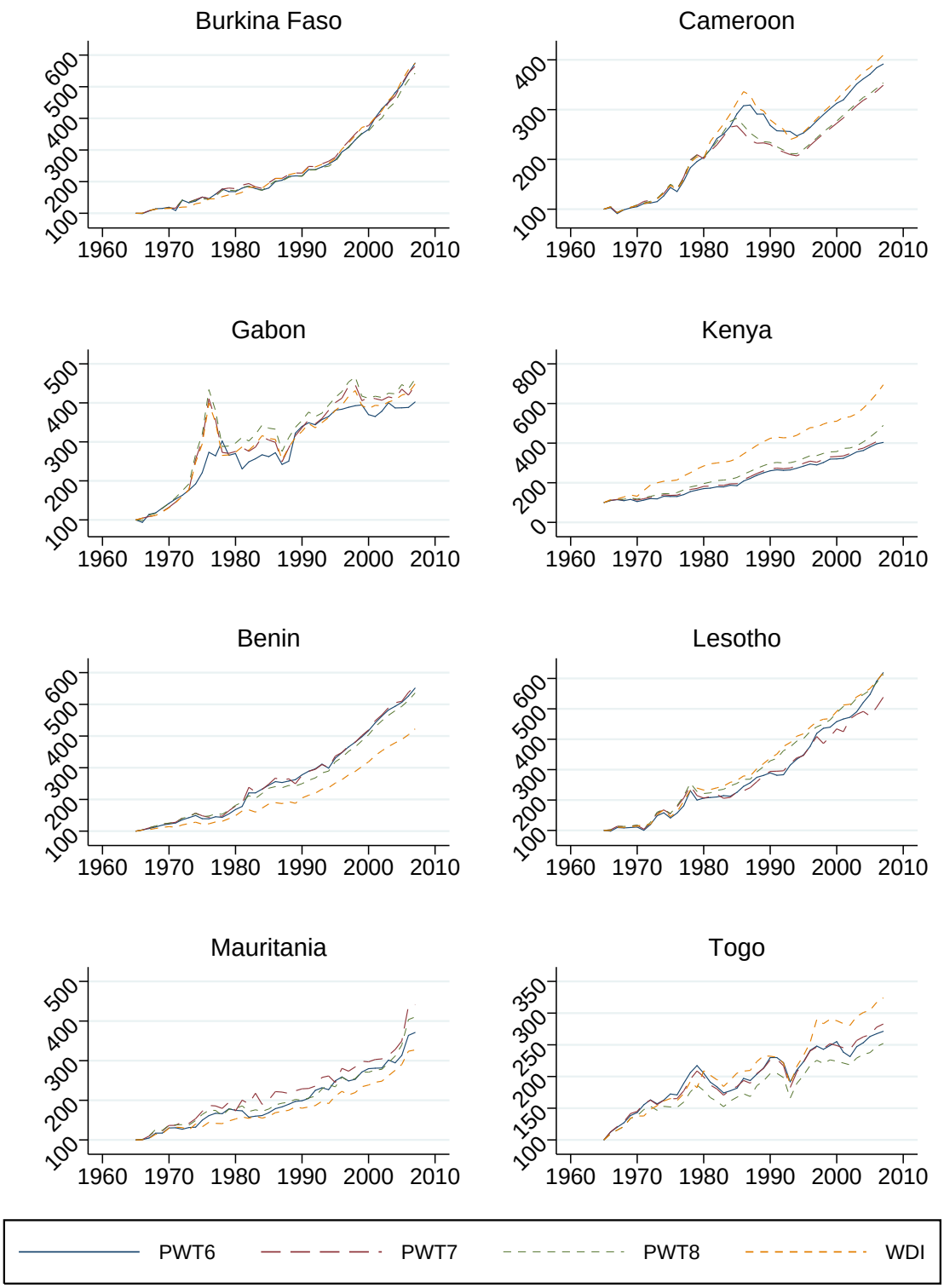
All GDP series in figure 1 are normalised to their respective 1965 levels in order to abstract from persistent differences in levels, emphasising growth rates instead. Only in one country, Burkina Faso, GDP takes an almost identical trajectory in all four datasets. In other countries, the discrepancies look rather well behaved, that is, they occur only at specific points in time or concern only a single series. Kenya is such a case, where WDI

[4]RGDPna is widely unaffected by some of the more fundamental innovations in PWT8, which are reflected in the newly introduced series RGDPe and RGDPo, and aim at rendering GDP figures comparable across both time and space. The authors do however recommend RGDPna for single country analyses mainly concerned with growth rates (Feenstra et al., 2015).

[5]RGDPL used to be constructed from the growth rates of individual expenditure shares, with the result that changes in the underlying prices could induce major differences in overall GDP growth rates. Section 3.3 discusses this issue in some more detail.

[6]While the coverage is larger for the GDP series as such, the corresponding expenditure shares are often missing.

Figure 1: GDP series from four sources, normalised to 1965 levels



indicates that GDP has increased by a factor of about 6.5 from 1965 to 2007, whereas the PWT measures agree on a factor of about 4. The opposite is the case in Benin, where WDI indicates a persistently lower growth rate subject to similar fluctuations as that reported by the PWT measures. In Cameroon, all measures follow an almost identical path until the mid-80's, but then split up: PWT8 and WDI register continued growth until the late 80's, followed by about a decade of recession, PWT6 and PWT7 start indicating a similarly severe recession earlier. The resulting differences are preserved in subsequent periods, where the datasets generally agree on the growth rates, but at different levels of (normalised) GDP.

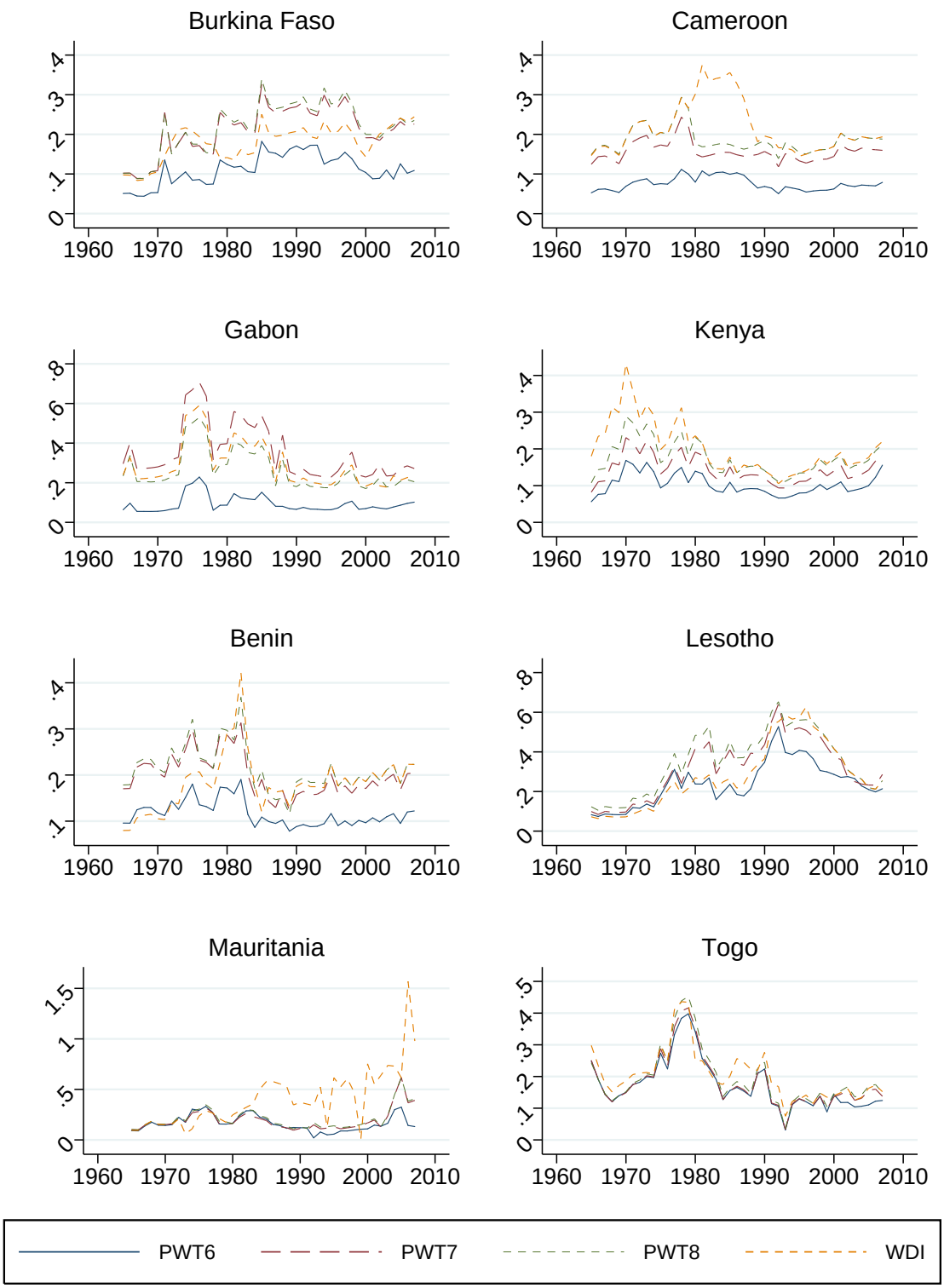
In Togo, Mauritania, Gabon and Lesotho the differences are perhaps the most striking and least tractable. While the general patterns tend to be the same (they agree on major booms and recessions), differences arise throughout the period without following an apparent pattern, leading the graphs to intersect sometimes multiple times. That is, none of the datasets systematically under- or over-reports growth, but the sign of their relative bias varies over time. The most striking single discrepancy may be the one between PWT6 and the remaining series in the mid-70's in Gabon. This is almost entirely explained by different underlying prices of investment, as will be discussed in more detail in section 3.3.

It is worth emphasising that there is no obvious pattern describing how the datasets behave relative to each other: In Gabon, PWT6 is the obvious outlier, in Cameroon (and in Lesotho, to some extent), PWT6 and PWT7 take one path, PWT8 and WDI the other, in Togo, PWT8 and WDI are precisely the ones the diverge most relative to each other. Only Kenya and Benin are consistent with the perhaps most intuitive expectation, that is, WDI diverging from otherwise consistent PWT measures. However, even in these two countries the divergence has opposite signs: in Kenya, WDI indicates higher growth, in Benin, it indicates lower growth than the remaining series.

In order to illustrate not only the divergence between GDP levels, but also in the relative movements of its components (which is the more influential aspect in the context of the study at hand), figure 2 plots the share of investment in GDP over the same period and for the same countries as those in figure 1. Note that the y-axis is scaled in order to depict a maximum of detail in the variation for each country. Its range is therefore informative in itself, and can vary with two factors: The temporal variation of the investment share within countries, and the discord between datasets regarding the investment share. For instance, investment in Lesotho varies from approximately 10% to more than 60% of GDP, but this is mainly due to temporal variation of similar amplitude in all datasets. In Cameroon, the scale is mainly stretched by upwards outliers in the 1980's in WDI (up to almost 40%), and consistently much lower estimates of less than 10% in PWT6. In most instances, the discrepancy is relatively constant over time, reflected in more or less parallel paths of the graphs; this applies in particular to Gabon, Kenya, and in a less pronounced manner to Burkina Faso, Togo, and Benin. The obvious outlier in the panel is Mauritania: While the PWT series already diverge substantially, this is dwarfed by the path suggested by WDI. Consistent with the other series until the early 1980's, the share of investment then skyrockets to levels of around 60% while the others agree on 10–20%, and reaches a peak of more than 150% of GDP (possible through a large trade deficit) where the PWT series report values between 20% and 40%.

Unlike in the GDP series depicted in figure 1, there seems to be a clear pattern across datasets in the investment share series: PWT6 almost consistently reports a lower share of GDP than the other datasets. This is mainly because PWT6 relies on price estimates

Figure 2: Shares of investment in GDP from four sources



from the ICP 1996, whereas the other series rely on those from 2005. The price estimates for investment are systematically lower when ICP 2005 prices are used, mechanically increasing the estimated real investment shares. Whether this reflects actual changes in prices, improved accuracy, or is a methodological artefact, is discussed in detail by Breton (2015) and Breton and García (2016), summarised in section 2.

Table 1: Correlations between growth rates of the core variables

| | | PWT6 | PWT7 | PWT8 |
|---|---|---|---|---|
| | PWT7 | 0.76 | | |
| Δ GDP | PWT8 | 0.75 | 0.87 | |
| | WDI | 0.65 | 0.75 | 0.85 |
| | PWT7 | 0.59 | | |
| Δ Investment | PWT8 | 0.72 | 0.76 | |
| | WDI | 0.04 | 0.05 | 0.07 |
| | PWT7 | 0.74 | | |
| Δ Consumption | PWT8 | 0.50 | 0.61 | |
| | WDI | 0.34 | 0.44 | 0.20 |
| | PWT7 | 0.65 | | |
| Δ Government | PWT8 | 0.65 | 0.96 | |
| | WDI | 0.51 | 0.75 | 0.75 |

Notes: The reported values are pairwise correlations between each of the variables in the four datasets employed. All correlations are significant at conventional levels, with the exception of those involving the WDI investment series. See main text for discussion.

Source: Author's calculations.

Table 1 reports the pair-wise correlations between the annual GDP growth rates, and the shares of investment, consumption and government expenditure across the four datasets employed in JMT's sample of countries in the analysis over the period from 1965 to 2007, subject to data availability. Given that they are conceptually in principle the same, it is a striking result that the *highest* correlation of the growth rates (first panel) between any of the datasets is 0.87 (PWT7 and PWT8), and goes as low as 0.65 (PWT6 and WDI). As much of the further analysis will focus on comparing results obtained from PWT6 (being JMT's original dataset) with those obtained from each of the other measures, it is also important to note that for PWT8 and WDI, the lowest correlation is precisely that with PWT6, and PWT7 is only marginally less correlated with WDI than with PWT6.

Looking at the second panel in table 1, the share of investment is only moderately correlated between PWT6 and PWT7 (0.59), and slightly more between PWT6 and PWT8 (0.72). Strikingly, seen from this perspective, the investment series in WDI shows almost no relationship with the PWT measures: all correlation coefficients are below 0.1, and none of them is statistically significant at the 5% level (the smallest $p$-value is 0.06 for WDI/PWT8). However, it needs to be pointed out that this is strongly driven by the case of Mauritania, where the WDI investment series obviously bears almost no relationship with the remaining ones (figure 2). When excluding Mauritania, the correlations with investment in WDI become: 0.46 (PWT6), 0.51 (PWT7), 0.63 (PWT8).

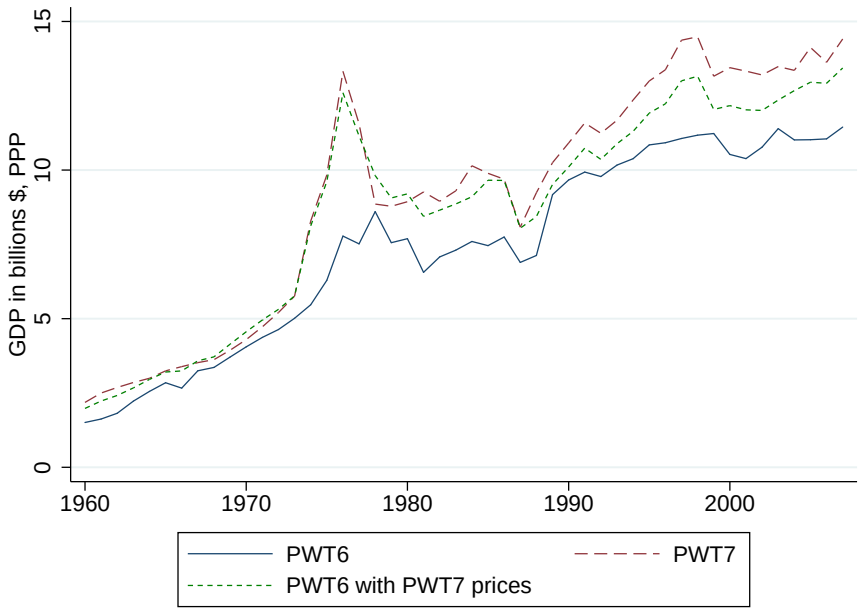
The third and fourth panel of table 1 report the correlations for the remaining two measures, the growth in consumption and government expenditure. The magnitudes of these coefficients are more or less in line with those observed for investment. Note however that WDI is now less of an outlier, although it still features the poorest correlation with any series (at its worst, 0.2 with PWT8 in the consumption series).

A general and intuitive pattern (although not universal) in table 1 is that the correlations go down as we move away from the 'original' dataset – PWT6, by these measures, is more similar to PWT7 than to PWT8, and the least similar to WDI, which is both the most recent dataset and comes from a different data provider.

## 3.3 The relative importance of revisions to price and NA data

Some major differences between the datasets can be explained with differences in the underlying prices and PPP estimates. To illustrate this, figure 3 plots the constant price real GDP series (RGDPL) of Gabon as reported by PWT6 and PWT7. While the series generally seem to take similar paths, there are periods of striking divergence. The most obvious one is perhaps in the mid-1970's, where PWT7 reports an enormous increase in GDP that is almost completely missing in PWT6. At its peak in 1976, PWT7 reports a GDP figure 71% higher than PWT6. Although the series do converge slightly in the aftermath and then evolve almost in parallel, PWT7 consistently reports a higher GDP than PWT6, on average by 23%.

Figure 3: GDP of Gabon according to PWT6 and PWT7, 1960-2007




Notes: The blue and the red line are the original RGDPL series from PWT6 and PWT7 respectively. The blue line is based on NA data and the methodology of PWT6, with the only difference that PPPs are constructed using PWT7 price data, reducing discrepancies dramatically. All series in 2005 $ (PPP).

This changes as I recompute the GDP series using the methodology and the national accounts data underlying PWT6, but the prices underlying PWT7. As discussed above, the latter are derived from the ICP 2005, while PWT6 is based on ICP 1996 prices (see appendix A.1 for the details of the computation). The resulting series is plotted in green

in figure 3: PWT6 and PWT7 are reconciled to a large extent, with a difference of less than 6% in 1976, compared to 71% before. The average difference between the series after 1976 drops from 23% to about 6%.

In fact, a closer look at the expenditure category prices in 2005 in Gabon reveals that the divergence between PWT6 and PWT7 almost entirely stems from the underlying price of investment (consistent with the findings of Breton and García (2016) discussed in section 2). While all other prices are approximately the same, investment is estimated to cost 193% of the price in the US in 2005 (which is the reference point) in PWT6, but only 51% in PWT7. As a result, a nominally important increase in investment expenditures in the mid-1970's is highly *de*flated in real terms in PWT6, but *in*flated in PWT7. In PWT7, the increase in real investment that enters GDP is consequently estimated to be almost four times higher. When repeating this exercise for all 36 countries in the sample, PWT6 and PWT7 can be largely reconciled in almost all countries. The exceptions are Ghana, Liberia, Somalia and the Seychelles, where differences are mostly due to revisions to the NA data. The relevant plots and a more detailed discussion are provided in appendix A.2.

In order to assess the relative importance of changes in the national accounts data and prices more formally, I construct a counter-factual GDP series that is based on PWT7 NA data, but PWT6 price estimates for every country in the sample. Table 2 summarizes the extent to which each series diverges from the original PWT7 in terms of levels and growth rates. The divergence is quantified by their mean *absolute* deviation (MAD) and mean deviation (MD). MAD averages over the absolute differences between the series irrespective of their sign, and is therefore a good measure of general divergence between the series. In MD, positive and negative differences can cancel each other out. It will therefore generally be smaller in magnitude, while its sign is indicative of the general direction of the divergence.

Table 2: Mean deviation and mean absolute deviation from PWT7

| | | Levels | Growth rates |
|---|---|---|---|
| PWT6 | MAD | 67.6% | 3.3% |
| | MD | 62.5% | 0.1% |
| PWT6 NA, PWT7 prices | MAD | 11.9% | 2.3% |
| | MD | -6.3% | 0.0% |
| PWT6 prices, PWT7 NA | MAD | 59.3% | 2.4% |
| | MD | 57.0% | 0.0% |

As suggested by the visual inspection of the graphs, the prices play a much bigger role in explaining the divergence in levels between PWT6 and PWT7 than changes in the NA data. PWT6 differs, on average, by 67.6% from PWT7 in our sample - almost all deviations being positive, as indicated by the MD. When keeping the underlying NA data constant but applying PWT7 prices, this difference shrinks to 11.9%. The underlying NA data itself contributes much less to the divergence in terms of levels: When using PWT6 prices with PWT7 NA data, the difference to the PWT7 series remains high at 59.3% difference on average – a relatively small improvement over the original PWT6 series.

Since inference in times series studies is based on dynamics within a country over time, we are particularly interested in the growth rates. Consider the second column of

table 2: The MAD between PWT6 and PWT7 is 3.3%. Compared to average growth rates reported in our sample of 36 countries between 1960 and 2007 of 3.7% (PWT6) or 3.8% (PWT7), this is a huge discrepancy of, on average, about 87% in any period. Note also that this divergence is not directed: The average growth rates are hardly different between PWT6 and PWT7: on average, PWT6 only indicates just about 0.1% higher growth rates. As opposed to the differences in levels, the underlying NA data and the prices have very similar impacts when it comes to growth rates: Reconstructing PWT6 using PWT7 prices reduces the MAD to 2.3%, using PWT6 prices but PWT7 NA data reduces it to 2.4%. Both changes in the underlying NA data and changes in the price series can therefore account for about one third if the differences in GDP growth rates between PWT6 and PWT7.

# 4 Baseline Study: Juselius, Møller and Tarp (2014)

Before proceeding to the robustness exercises, I will briefly outline the methodology of JMT, which I adopt. Section 4.1 describes the general CVAR framework, and section 4.2 describes the way in which JMT draw inference concerning the long-run effectiveness of aid.

## 4.1 The Cointegrated VAR Framework

JMT's paper analyses the long-run effect of foreign aid on key macroeconomic variables in 36 sub-Saharan African countries, specifying a Cointegrated VAR (CVAR) model for each country individually. While these models differ in terms of lag-length, deterministic components and rank restrictions, they share the basic structure that can be captured in the following moving average (MA) representation:

$$\mathbf{X}_t = \mathbf{C}\sum_{i=1}^{t}\varepsilon_i + \mathbf{C}\,\Phi\sum_{i=1}^{t}\mathbf{D}_i + \mathbf{C}^*(L)\varepsilon_t + \mathbf{C}^*\,\Phi\,\mathbf{D}_t + \mathbf{A}_0 \tag{1}$$

where $\mathbf{X}_t$ is a vector of $p = 5$ dependent variables, $\mathbf{X}_t = [aid_t,\ y_t,\ inv_t,\ c_t,\ g_t]'$ (inflow of foreign aid, total GDP, investment, private consumption and government expenditure in year $t$). $\mathbf{C}$ is a $p \times p$ matrix of rank $p-r$, $r$ being the number of cointegrating relations between the variables. $\mathbf{D}_i$ ($\mathbf{D}_t$) are $m \times 1$ vectors containing the $m$ deterministic components of the model at time $i$ ($t$), such as trends in the variables or dummies accounting for extraordinary events. These enter the model weighted by the coefficients in the $p \times m$ matrix $\mathbf{\Phi}$. $\mathbf{C}^*(L)$ is a stationary lag polynomial and $\mathbf{A}_0$ contains the initial values of the variables in $\mathbf{X}_t$ and the initial values of the short-term dynamics. The interested reader may refer to JMT (p. 7-11), who lay out their methodology in more detail.

In the present analysis, the focus is on the long-run impact matrix $\mathbf{C}$, which has the following structure:

$$\mathbf{C} \quad = \quad \begin{array}{c} \\ aid_t \\ y_t \\ inv_t \\ c_t \\ g_t \end{array} \begin{array}{c} \begin{array}{ccccc} \hat{\varepsilon}_{aid} & \hat{\varepsilon}_y & \hat{\varepsilon}_{inv} & \hat{\varepsilon}_c & \hat{\varepsilon}_g \end{array} \\ \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \\ c_{21} & c_{22} & c_{23} & c_{24} & c_{25} \\ c_{31} & c_{32} & c_{33} & c_{34} & c_{35} \\ c_{41} & c_{42} & c_{43} & c_{44} & c_{45} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} \end{pmatrix} \end{array} = \begin{pmatrix} c_{11} & \mathbf{C_{12}} \\ \mathbf{C_{21}} & \mathbf{C_{22}} \end{pmatrix} \tag{2}$$

The element $c_{ij}$ in row $i$ and column $j$ describes the impact that the cumulated exogenous shocks, measured as the residuals of the respective equation, on variable $j$ have exerted on variable $i$ in the long run. In this sense, for example, $c_{21}$ can be interpreted as the long run effect of exogenous shocks to foreign aid on GDP.

In line with JMT, equation (2) provides a convenient decomposition of the matrix $\mathbf{C}$. The four coefficients describing the long-run impact of aid on the other variables are contained in $\mathbf{C}_{21}$. It follows from JMT's research question—the long-run effectiveness of foreign aid—that this will be at the centre of my discussion. $\mathbf{C}_{12}$ contains the long-run effects of the other macrovariables on aid, and is therefore informative when asking questions about the potential (and plausible) endogeneity of aid. This will however not be the subject of the study at hand. The $(p-1) \times (p-1) = 4 \times 4$ matrix $\mathbf{C}_{22}$ contains the remaining parameters, having analogous interpretations for the interaction between the macrovariables variables other than aid.

### 4.2 Inference in JMT

JMT's answer to the main question, the long-run impact of foreign aid on economic growth, is based on the sign and magnitude of the $t$-ratios associated with the coefficients of the vector $\mathbf{C}_{21}$. Aid is considered effective if either the coefficient of $y_t$, or $inv_t$, is positive and significant (absolute $t$-ratio larger than 2), or if both are. While the representation of the results also highlights marginally insignificant coefficients (absolute $t$-ratio between 1.6 and 2), these do not impact on the final inference. Aid harmfulness is defined analogously, with opposite signs. As their focus is on long-term growth, JMT do not consider the coefficients of $c_t$ and $g_t$ for their final inference. They do, however, discuss and report them, and I will consider them in the subsequent robustness checks.

In order to avoid a bias resulting from the researchers' economic prior, JMT consider the results from two different angles, adopting both the prior of effectiveness and that of harmfulness. The critical element here is the choice of the cointegration rank $r$, which especially in relatively short samples is not always clear cut and can have substantial impact on the inference. Instead of only determining one preferred choice of rank, JMT determine a best and second best choice (see also section 5.3.1). Under the prior of aid effectiveness, they report the preferred choice of rank if it indicates effectiveness. If the preferred choice of rank does not indicate effectiveness, but the second best choice does, they report the latter. The same procedure is then applied under the prior of harmfulness, but looking for negative coefficients.[7]

## 5 Assessing the impact of data discrepancies

In order to assess the impact of differences between the datasets, I proceed in two steps. In the first step, I apply precisely the statistical models derived by JMT to the alternative datasets. As these models have been tailored to the PWT6 data originally, it may however be the case that, when confronted to the new data, these models are technically misspecified, compromising the validity of any inference. In a second step, I therefore proceed to re-specifying the statistical models for each dataset individually, applying the same statistical criteria as JMT, in the 12 countries for which data is available from all

[7] JMT also report results using strictly the preferred rank, and first, second and third best choice of rank. Their results are largely robust to these alternative search algorithms.

four sources (PWT6, PWT7, PWT8, WDI). Before engaging in these exercises, I establish the criteria by which I will assess the robustness of the results.

## 5.1 Criteria to assess the stability of the results

Especially since JMT's mode of inference is rather unconventional, it is necessary to establish some sensible and objective criteria prior to assessing the stability of their results. My discussion will evolve around:

(i) The number of qualitative changes (that is, sign and/or significance) in *any* of the coefficients, i.e. those of $y_t$, $inv_t$, $c_t$ and $g_t$. While JMT do not base their overall inference on $c_t$ and $g_t$, they are still reported and discussed.

(ii) The number of qualitative changes in the *most relevant* coefficients, $y_t$ and $inv_t$.

(iii) The number of qualitative changes in the *inference by country*. This is not equivalent to the previous point, as the inference is a joint product of the two coefficients of $y_t$ and $inv_t$ - even if both change sign/significance, we would still infer effectiveness as long as one of them is positive significant.

(iv) Any changes in the *overall conclusion*. JMT count 27 out of 36 cases of aid effectiveness. Irrespective of the results obtained for the previous criteria, this ratio may change or remain approximately constant, as reversals in one country may be compensated for by reversals in the opposite direction in other countries.

For criteria (i) and (ii) (coefficient specific), I consider changes in the coefficients a *reversal* if their absolute $t$-value changes at least from $|t| > 2$ to $|t| < 1.6$ (loses significance), or does the opposite (gains significance). The same is true if the coefficient changes signs, unless it is insignificant in both instances (that is, in JMT's original study and in my replication).

Table 3 provides a full summary of this classification. The notation follows JMT: a plus $(+)$ corresponds to a positive $t$-ratio larger than 2, meaning that the variable was positively affected by aid in the long run, while minus $(-)$ indicates a $t$-ratio smaller than $-2$, indicating a negative relationship. The subscript 0 indicates absolute $t$-ratios between 2 and 1.6 (marginally significant), the subscript 00 an absolute $t$-ratio smaller than 1.6 (insignificant). I label coefficients *inconclusive* if the $t$-ratio remains of the same sign, and changes from fully insignificant ($|t| < 1.6$) to marginally significant ($2 > |t| > 1.6$), from marginally significant to significant ($|t| > 2$), or does any of these changes in the opposite direction.

Regarding criterion (iii), the inference by country is well defined in the original paper and takes three possible values (effectiveness, harmfulness, or insignificant), and I classify any change from or to insignificance as inconclusive, and a change from harmfulness to effectiveness or vice versa as a reversal.

Overall inference, criterion (iv), boils down to a simple ratio between the number of cases of effectiveness and harmfulness respectively, and does not require further definition.

Note also that JMT's approach of adopting different economic priors (see section 4.2) adds a further complication: Cases can arise where, while the inference for a country (criterion (iii)) does not change, it will be based on a different choice of rank, with the originally reported rank now indicating a different result. I will count the rare cases where this occurs as consistent if the inference remains the same, no matter the rank this inference is based on.

Table 3: Classification of the replication results

| | | JMT | | | | | |
|---|---|---|---|---|---|---|---|
| | | $-$ | $-_0$ | $-_{00}$ | $+_{00}$ | $+_0$ | $+$ |
| Replication | $-$ | ✓ | · | ✗ | ✗ | ✗ | ✗ |
| | $-_0$ | · | ✓ | · | · | ✗ | ✗ |
| | $-_{00}$ | ✗ | · | ✓ | ✓ | · | ✗ |
| | $+_{00}$ | ✗ | · | ✓ | ✓ | · | ✗ |
| | $+_0$ | ✗ | ✗ | · | · | ✓ | · |
| | $+$ | ✗ | ✗ | ✗ | ✗ | · | ✓ |

Legend: ✓= Consistent; · = Inconclusive; ✗ = Reversal. + indicates a positive coefficient, − a negative one. The subscript '0' indicates an absolute $t$-ratio between 1.6 and 2, '00' indicates one lower than 1.6. The absence of a subscript indicates an absolute $t$-ratio $> 2$.

## 5.2 Simple replication

The first replication uses the models developed in JMT, reported in table 2 of their paper. This involves a replication using the original data (PWT6, except for Sudan where JMT use WDI data) for all 36 countries, and the same exercise employing the alternative datasets. JMT's mode of inference of considering the first best and second best choice of rank requires two estimations per country and dataset (one for each choice of rank). Besides the original data for the 36 countries in JMT's sample, we have sufficient data for 36 (PWT7), 33 (PWT8), and 13 (WDI) of these countries in the alternative datasets, each of which has to be estimated twice, leaving us with 236 country-, dataset-, and rank-specific estimations. These have been carried out in an appropriately modified version of CATS in RATS version 2.6.[8]

### 5.2.1 Consistency across datasets

Table 4 summarises the key results from this first exercise. The four panels correspond to criteria (i)–(iv) as described above. The first two columns in table 4 report the results from the immediate replication of JMT, using the same data as them except for some possible revisions to the DAC aid data. Throughout the criteria, it is apparent that these revisions only have a very limited impact on the results: 127 out of 144 coefficients remain completely identical, and only 7 change in a significant way, that is, they are counted as reversals (panel (i)). Similar figures apply for the core coefficients, those of GDP and investment (panel (ii)). In all but one country (Mauritania), the final inference remains the same (panel (iii)). In Mauritania, the conclusion changes from effectiveness to harmfulness, leading us to conclude effectiveness in 26 countries under the prior of aid effectiveness, compared to 27 in JMT's original results. Under the prior of harmfulness, nothing changes and evidence for this hypothesis can be found in 10 countries.

Using alternative datasets has a sizeable impact on the results: The ratio of consistent coefficients (panel (i)) is just under two thirds in all alternative datasets (58%–63%), and about one quarter of them are reversed (23%–28%). Our main coefficients (panel (ii)) are

[8]Replication programs and data can be obtained at `http://bit.ly/2sVH63e`. Parts of the relevant code are embedded in proprietary program files and can therefore not readily be shared, but detailed documentation on the necessary modifications can be provided.

Table 4: Summary of replication results

| | | Replication | | Alternative Datasets | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PWT6 | | PWT7 | | PWT8 | | WDI | |
| **(i)** | **Any coefficient** | | | | | | | | |
| | Consistent | 127 | *88%* | 89 | *62%* | 76 | *58%* | 33 | *63%* |
| | Inconclusive | 10 | *7%* | 15 | *10%* | 22 | *17%* | 7 | *13%* |
| | Reversal | 7 | *5%* | 40 | *28%* | 34 | *26%* | 12 | *23%* |
| **(ii)** | **Main coefficients** | | | | | | | | |
| | $y_t$ | | | | | | | | |
| | Consistent | 33 | *92%* | 20 | *56%* | 21 | *64%* | 9 | *69%* |
| | Inconclusive | 2 | *6%* | 8 | *22%* | 5 | *15%* | 0 | *0%* |
| | Reversal | 1 | *3%* | 8 | *22%* | 7 | *21%* | 4 | *31%* |
| | $inv_t$ | | | | | | | | |
| | Consistent | 32 | *89%* | 20 | *56%* | 16 | *48%* | 9 | *69%* |
| | Inconclusive | 3 | *8%* | 3 | *8%* | 6 | *18%* | 3 | *23%* |
| | Reversal | 1 | *3%* | 13 | *36%* | 11 | *33%* | 1 | *8%* |
| | $y_t$ *or* $inv_t$ | | | | | | | | |
| | Consistent | 29 | *81%* | 13 | *31%* | 12 | *33%* | 7 | *54%* |
| | Inconclusive | 5 | *14%* | 11 | *26%* | 10 | *28%* | 2 | *15%* |
| | Reversal | 2 | *6%* | 18 | *43%* | 14 | *39%* | 4 | *31%* |
| **(iii)** | **Country-wise inference** | | | | | | | | |
| | Consistent | 35 | *97%* | 24 | *67%* | 20 | *61%* | 10 | *77%* |
| | Inconclusive | 0 | *0%* | 11 | *31%* | 11 | *33%* | 3 | *23%* |
| | Reversal | 1 | *3%* | 1 | *3%* | 2 | *6%* | 0 | *0%* |
| **(iv)** | **Overall conclusion** | | | | | | | | |
| | *Prior: Effectiveness* | | | | | | | | |
| | Effective | 26 | | 18 | | 13 | | 6 | |
| | Insignificant | *8* | | *17* | | *17* | | *7* | |
| | *Prior: Harmfulness* | | | | | | | | |
| | Harmful | 10 | | 9 | | 7 | | 3 | |
| | Insignificant | *18* | | *20* | | *21* | | *7* | |
| | **Sample** | **36** | | **36** | | **33** | | **13** | |

Notes: The table reports criteria (i)-(iv) as described in section 5.1 for all datasets when using the models as derived by JMT. The first column or each dataset reports the absolute number of coefficients (criteria (i) and (ii)) or countries (criteria (iii) and (iv)), the second column the respective share of the total number of coefficients or countries in the respective set-up.

Source: Author's calculations.

no exception to this pattern. The last three rows of panel (ii) count whether either of the two main coefficients has changed sign or significance. In PWT7 and PWT8, these coefficients remained jointly stable in only 31% and 33% of all countries respectively, while at least one of the coefficients was subject to a reversal in 43% and 39% of the countries in the sample. WDI performed substantially better in this respect, with 54% of all countries having seen no changes in either of these coefficients.

The inference by country (panel (iii)) does not need to change in all of these cases, as a change in the coefficient of GDP can be compensated by a stable (or inconsistent in the opposite direction) coefficient of investment, and vice versa. In PWT7 and PWT8, the inference by country remains identical in 67% and 61% of cases, and gets reversed in 3% (1 case) and 6% (2 cases) respectively. Similar to what we saw for criterion 2, WDI yields results that are more consistent with JMT – inference is identical in 77% of the countries in the reduced sample (10/13) and is reversed nowhere.

Regardless of the relatively large variation in individual coefficients and country-wise inference, the overall conclusion of aid effectiveness (panel (iv)) remains unchallenged throughout all datasets; it is, however, slightly weakened. The count of cases of effectiveness versus cases of harmfulness (under the respective priors), goes down from 26:10 in my replication of JMT using PWT6 (27:10 in the original study) to 18:9 in PWT7, 13:7 in PWT8, and 6:3 in WDI. While for every case of harmfulness I (JMT) find 2.6 (2.7) cases of effectiveness using PWT6, there are about 2.0 cases of effectiveness for each case of harmfulness in the alternative datasets.

Note that while the results for WDI are proportionally speaking more in line with JMT than those obtained with PWT7 and PWT8, the significantly smaller sub-sample contained in WDI consists of countries that also exhibit more stable results in the other datasets. In the sub-sample of countries that are not covered in WDI, PWT7 and PWT8 yield consistent results only in 56% of the countries, compared to 77% in the sub-sample covered by WDI.[9] This could be reflecting a selection bias as a function of the quality of the original data, if for instance WDI were more conservative in 'filling the gaps'.

### 5.2.2 Consistency across countries

The share of consistent coefficients, excluding the PWT6 replication, ranges from 0% in Botswana and Lesotho, to 100% in Chad, Burkina Faso, Cameroon, Kenya and Gabon. The median proportion of consistent coefficients is 50%, the mean 57%. A table summarising the consistency by coefficients for each country can be found in appendix B.

In about half of the sample, 17 countries, the inference remains stable throughout all datasets. On the other hand, 9 countries do not yield the same inference with any of the new data. Only in 3 out of a total of 82 cases (given sample sizes of 36, 33 and 13 in the alternative datasets), the inference is reversed. These are Lesotho in PWT7 and PWT8, and Liberia in PWT8; all three reversals correspond to a switch from effectiveness to harmfulness. All other changes in country-wise inference correspond to a loss of significance.

Note that it is difficult to predict whether the results for a country will be stable with respect to changes in the data or not. As documented in appendix C, a variety of measures of divergence (in levels, shares, relative shares) or data quality were only very weakly associated with the robustness of the results. In fact, the initial results seemed

[9]The fact that this corresponds exactly to the ratio found in the replication using WDI is coincidental; only half of the reversed inferences are associated with the same countries as in WDI.

to be better predictors of their own consistency than any properties of the data: Initially insignificant coefficients were much more likely to remain insignificant in the replications than significant ones to keep their sign and significance. For the 8 countries that had no single significant coefficient in the original results, 80% of the coefficients were also insignificant in the replications. Within all other countries, 50% kept their sign and significance.

## 5.3 Re-specification for selected countries

The second exercise aims at exploring the full potential impact of the data on the results. It takes into account a fundamental aspect of the philosophy underlying the CVAR approach (and multivariate time series analysis more generally), namely that of 'allowing the data to speak freely' (Hoover et al., 2008, title): The priority lays in creating a statistically adequate representation of the data, avoiding strong theoretical assumptions *a priori*. In this spirit, each country-specific model developed by JMT, and thus applied in the previous section, has been specified as a function of the underlying data, that is PWT6 and the DAC ODA data. Given the sometimes striking divergence between the datasets, it seems reasonable to expect that differences may emerge not only in the results within the same models, but in the models themselves. I will therefore re-specify the models for the sub-sample on 12 countries where data is available for all datasets (the subset of countries covered by WDI bar Sudan, where JMT used WDI).

A priori, it is unclear what to expect from the results of this exercise. On the one hand, it might be the case that the changes observed in the previous exercise were partly a result of the models being effectively misspecified for the alternative datasets. In this case, it is possible that the data do indeed tell the same story, once analysed with adequately re-specified models. On the other hand, it may be the case that using the exact same models as JMT actually obscured the true impact of the changes in the data. Adjusting the models to the new data may then amplify the changes between datasets, and in fact induce even larger changes in the results.

### 5.3.1 Model Specification Procedure

The models employed in JMT have the following variable elements: deterministic components such as trends and dummies, the lag-length, and the cointegration rank. While the econometric literature offers a plethora of formal criteria to specify either of these components, the sample sizes in this application ($T \approx 40$) undermine the power of some of the relevant tests. Furthermore, different criteria typically offer a slightly different angle at the data, and will sometimes indicate different choices. The resulting trade-offs eventually need to be resolved by the researcher's judgement. This section discusses the model specification procedure and reports the relevant criteria for each of the 48 models specified (4 datasets for 12 countries).

**Deterministic components** The choice of the deterministic components in this model consists of two elements, linear trends and dummy variables that account for extraordinary events or regime changes.

**Linear trends** Given the macroeconomic nature of our data, we would typically expect the presence of linear trends. These can be incorporated in the VAR in different

ways, and the decision essentially boils down to the question whether the trends cancel in the cointegrating relations or not. In the case where they do cancel out, the most accurate specification would include an unrestricted constant term in the equation. If they do not cancel out, a trend that is restricted to the cointegrating relations should be included; this is the case when either some of the variables in $\mathbf{X}_t$ or any of the cointegrating relations are trend-stationary. For a detailed discussion, see Juselius (2006, chapter 6). One way of approaching the issue is to tentatively include a trend that is restricted to the cointegrating relations, which can then be tested for significance. In all of the present cases, this evidence unambiguously points towards the inclusion of a trend restricted to the cointegrating relations (see Juselius, 2006, p. 100, Case 4).

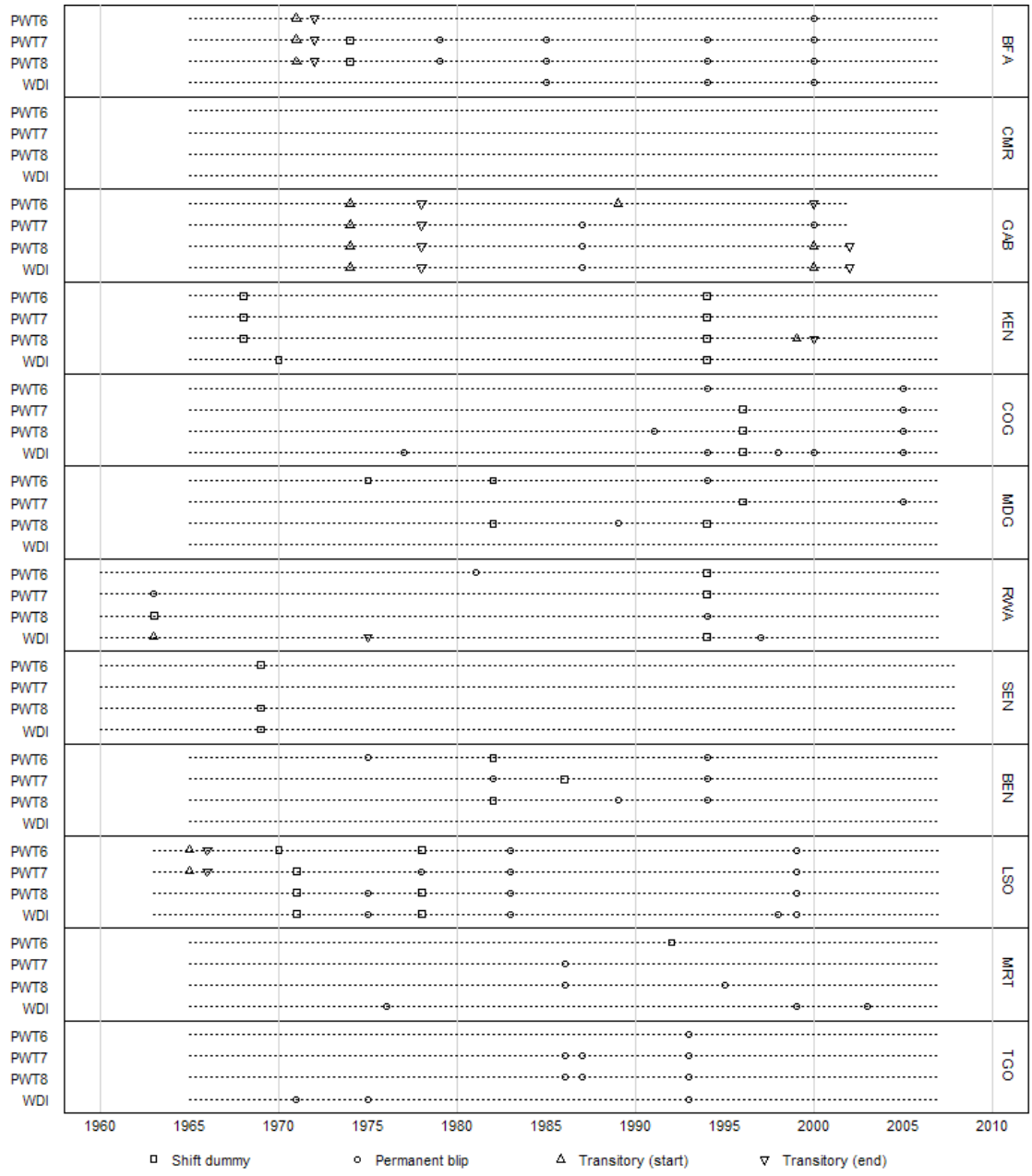
**Dummy variables** The more contentious, and less clearly defined choice, concerns the inclusion of dummy variables in order to account for extraordinary events such as droughts, floods, social unrest, or changes in equilibrium relations due to, e.g., regime changes. As it is difficult *a priori* to determine which historical events have an impact significant enough to enter the model, Juselius (2006, chapter 6.6) suggests to first scrutinise the data and the residuals from the baseline VAR, in order to determine where it is required to correct for outlier observations; the modelling does thus first depend on the statistical evidence, which is then complemented by institutional and historical knowledge.

In line with JMT and Juselius (2006), I use three classes of dummies: Permanent blip dummies, labelled $D_pZZ_t$, having the structure $[0, 0, 1, 0, ..., 0]$, transitory blip dummies $D_{tr}ZZ_t$ with the structure $[0, 0, 1, -1, 0, ..., 0]$, and shift dummies $D_sZZ_t$, restricted to the cointegrating relations and taking the form $[0, 0, 1, 1, 1, ..., 1]$, indicating a shift in the equilibrium mean of the cointegrating relations. From a data perspective, the indications for the inclusion of a dummy are primarily derived from an inspection of the residuals. Large residuals, 'large' here being construed as corresponding to approximately 3 standard deviations, indicate that there may be an extraordinary event which, if not accounted for appropriately, would distort the analysis.[10] A unique blip in the error series, for instance, if it is not reverted by a shock in the opposite direction in one of the following periods, may then indicate a unique event (e.g., a drought) that permanently affected the economy. A temporary blip, for example, a large positive residual followed by a large negative residual in the following period, is an indication for a transitory intervention; typical cases would include a period of expansive monetary policy, compensated for by contractionary policy later on, or temporary fiscal stimuli. The determination of a shift dummy, $D_sZZ$, is in practice less straightforward, as it accounts for level changes in the long-run equilibrium, which are not readily observable from looking at the residuals. However, graphical inspection of the cointegrating relations can provide some indication. Furthermore, Juselius (2006, chapter III.9) proposes a battery of recursive and backward recursive tests that can provide indications for shifts in the equilibrium relations, which are taken into consideration.

For the final choice of the dummy variables, the statistical evidence is further complemented with historical data from the UCDP/PRIO Armed Conflict Dataset (Pettersson and Wallensteen, 2015) and the EM-DAT International Disaster Database (Guha-Sapir

[10]The outliers detected this way were generally consistent with those discerned through the dummy saturation procedure implemented in Autometrics (see Doornik, 2009). Due to the fundamental importance of institutional knowledge the present set-up, careful inspection of the residual series generally served as the primary source of information.

Figure 4: Timeline of dummy variables



et al., 2014)[11], as well as knowledge about economically relevant historical episodes, most of which are documented in Ndulu et al. (2008). Figure 4 depicts the resulting dummy variables on timelines corresponding to the respective sample periods for each country and dataset, where the top line for each country corresponds to the model for the original PWT6 data, the second to PWT7, the third to PWT8 and the fourth to WDI. The presence of apparent bulks of dummy variables within each country across datasets illustrates the fact that the datasets tend to agree on the impact of major events. Take for instance $Dp83_t$ in Lesotho, the year of a severe drought affecting most of the population (Guha-Sapir et al., 2014), reflected here in a dramatic drop in investment. Another consistent outlier is $Dtr7478_t$ in Gabon, corresponding to a unusually large temporary increase in investment, likely spurred by sharp increases in the price of oil after the 1973 oil crisis. Or the $Ds94_t$ shift dummy in Kenya, which coincides with a severe drought coupled with a dysentery epidemic costing about 1000 lives (*ibid.*). The latter also illustrates the trade-off one faces when weighing institutional knowledge against the statistical evidence. While the historical events are themselves not necessarily susceptible to induce a shift in long-run equilibrium relationships, the statistical evidence suggests precisely this, with the recursive test for fluctuations of the eigenvalue indicating non-constancy of our $\alpha$ and $\beta$ parameters, with a sharp decline in the test statistic in 1994 for the third cointegrating relation. The inclusion of $Ds94_t$ establishes constancy over the entire sample period, justifying the inclusion both in JMT's and in my models.

On the other hand, there are about 25 dummy variables that are specific to the respective dataset, such as in Togo in 1971 (a decline in GDP and investment in the year of a drought affecting about 150.000 people (Guha-Sapir et al., 2014)) or 1975/76 (a spike in consumption and government spending coinciding with soaring phosphate prices, one of Togo's staple exports (Ndulu et al., 2008, Vol. 1, p. 150)). Particularly divergent, in terms of the modelling of extraordinary events, is Mauritania, where the 7 dummies included across the models occur in 6 different years, thus only agreeing once. Nevertheless, all of them can plausibly be attributed to historical events (1976 is the year of a severe drought effectively affecting the entire population; 1986 falls within a period of severe ethnic tensions and shortly follows a coup d'état by Ould Taya; 1992, a shift in equilibrium means, is the year of the adoption of a new constitution establishing the fourth republic; 1995, 1999 and 2004 have all seen major floods (Guha-Sapir et al., 2014; Seely, 2005)).

Note that some of the peculiarities in the data cannot that clearly be attributed to historical events, in which case the trade-off is between the statistical fit of the model, and its institutional/historical justifiability. The 1971/72 transitory dummy in Burkina Faso, indicated in all datasets except WDI, is such a case. While this falls within a period of some political instability, with a tightening of military control and a decline of the civilian role (Ndulu et al., 2008, Vol. 2, p.8 Appendix), there is no compelling historical evidence justifying the transitory nature of the shock. It is, however, by some margin the specification that best fits the pattern of the residuals, and my models therefore follow JMT in accounting for it with $Dtr71_t$. In other cases, the trade-off has been resolved in favour of institutional adequacy, for instance in the case of an outlier in 1978 in the PWT7 Cameroon model: The model achieves a reasonable fit overall without a correction, and I could not identify a strong case for an extraordinary event in the country's history.

Overall, JMT's and my models (consistent in PWT6 except for Burkina Faso, the

[11]The events documented in these two databases have been matched with the dummies included in all models in two tables available at `http://bit.ly/2tYKSu2`.

Republic of Congo and Madagascar, where the aid data appears to have significantly changed the requirements) follow the same patterns and account for the statistical and historical evidence in a coherent manner. To summarise, the data tends to tell a similar story – stylised in figure 4 – but puts varying emphasis on individual episodes.

**Choice of lag-length** Determining the optimal lag-length of the resulting VARs boils down to a trade-off between preserving a maximum of information, and retaining a reasonable number of parameters, especially given the relatively small number of ca. 40 observations for each country (certainly *small* in the context of a time series analysis). My choices are based on three groups of criteria, in line with Juselius (2006): (i) a likelihood test for lag reduction, (ii) information criteria, namely the Schwarz Criterion (SC) and the Hannan-Quinn Criterion (HQC), and (iii) a Lagrange-Multiplier test for no autocorrelation of the residuals. The $p$-values associated with (i) and (iii), as well as the values of the (ii) ICs are collected in table 5.

The (i) test for lag reduction tests for a significant decrease in the (log-) likelihood when adopting a more parsimonious model, which indicates a poorer fit of the latter. The first column of table 5 reports the outcome only of the test for the reduction to the final choice of lag-length $k^*$ from $k^* + 1$ in order to save space. Note that this is rejected in almost all cases, including those exactly replicating JMT's models (all PWT6 except Burkina Faso, Congo, and Madagascar), indicating that higher lags may still contain valuable information. Although not explicitly reported here, this test rejects almost any reduction up to at least the fourth lag, which was the highest lag included in the procedure, therefore indicating prohibitively high autoregressive orders given the small sample sizes we are facing. This is because it does not account for the fact that each new lag increases the number of parameters by $p^2$, thereby rapidly consuming the already scarce degrees of freedom.

The (ii) information criteria, explicitly designed to take this trade-off into account, will therefore tend to indicate shorter lag-lengths. Columns 3 to 6 of table 5 report the values of the SC and the HQC for both $k^*$ and $k^* + 1$; the lag length associated with the smaller value is considered more favourable by the respective criterion. In all cases but five, the ICs unambiguously indicate that $k^*$ is preferable over $k^* + 1$; and $k^* = 1$ in all cases but Rwanda PWT7 and PWT8, and Senegal PWT7, where $k^* = 2$. In Kenya, HQC favours a lag-length of $k = 2$ in PWT6, PWT7 and WDI. Similarly, the ICs disagree in Congo WDI, Rwanda PWT6, and Senegal PWT8. In these ambiguous cases, I opt for the more parsimonious specification. The fact that the ICs almost unambiguously indicate a short lag length is not surprising; both SC and HQC penalise the inclusion of new parameters more harshly, the shorter the T, and will therefore apply particularly high penalties in the short samples at hand.

Finally, columns 7-9 report the $p$-values from the (iii) LM test for no residual autocorrelation in the first lag of the VAR($k^*$) models, and in the first and second lag of the VAR($k^* + 1$) models. The null hypothesis of no autocorrelation is rejected at the 5% level only three times in the first lag in the VAR($k^*$) models (Kenya WDI, Mauritania PWT6, and Congo WDI). In most cases, it does not oppose $k^* + 1$ either, although the test indicates residual autocorrelation seven times in the first lag and four times in the second lag.

Overall, all datasets appear to be best described with a lag-length of $k = 1$, with the exception of Rwanda PWT7 and PWT8, and Senegal PWT7. The most contentious choices are arguably Kenya and Congo in WDI, where HQC favours a higher lag-length,

Table 5: Criteria for the choice of the lag-length

| | | Schwarz | | Hannan-Quinn | | Autocorrelation | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Data* | *LR* | $k^*$ | $k^*+1$ | $k^*$ | $k^*+1$ | $k^*,1$ | $k^*+1,1$ | $k^*+1,2$ | $k^*$ |
| **Burkina Faso** | | | | | | | | | |
| PWT6 | 0.00 | -22.44 | -21.39 | -23.66 | -23.28 | 0.47 | 0.74 | 0.39 | 1 |
| PWT7 | 0.00 | -24.57 | -23.65 | -26.46 | -26.21 | 0.98 | 0.91 | 0.14 | 1 |
| PWT8 | 0.00 | -24.42 | -23.46 | -26.31 | -26.02 | 0.97 | 0.85 | 0.09 | 1 |
| WDI | 0.00 | -22.72 | -21.87 | -24.07 | -23.89 | 0.24 | 0.02 | 0.41 | 1 |
| **Cameroon** | | | | | | | | | |
| PWT6 | 0.12 | -23.57 | -22.10 | -24.51 | -23.72 | 0.57 | 0.75 | 0.60 | 1 |
| PWT7 | 0.00 | -24.07 | -22.94 | -25.01 | -24.56 | 0.31 | 0.17 | 0.02 | 1 |
| PWT8 | 0.00 | -23.67 | -22.61 | -24.62 | -24.23 | 0.17 | 0.26 | 0.03 | 1 |
| WDI | 0.90 | -22.10 | -20.21 | -23.04 | -21.82 | 0.97 | 0.08 | 0.75 | 1 |
| **Gabon** | | | | | | | | | |
| PWT6 | 0.00 | -16.83 | -15.45 | -18.43 | -18.07 | 0.40 | 0.42 | 0.17 | 1 |
| PWT7 | 0.00 | -17.72 | -16.20 | -19.61 | -19.25 | 0.56 | 0.72 | 0.21 | 1 |
| PWT8 | 0.00 | -17.92 | -16.45 | -19.81 | -19.51 | 0.56 | 0.53 | 0.13 | 1 |
| WDI | 0.00 | -17.88 | -16.50 | -19.77 | -19.55 | 0.43 | 0.54 | 0.34 | 1 |
| **Kenya** | | | | | | | | | |
| PWT6 | 0.00 | -23.88 | -23.21 | -25.36 | -25.36 | 0.15 | 0.61 | 0.76 | 1 |
| PWT7 | 0.00 | -24.82 | -24.21 | -26.30 | -26.37 | 0.09 | 0.52 | 0.32 | 1 |
| PWT8 | 0.00 | -24.95 | -24.18 | -26.57 | -26.47 | 0.20 | 0.04 | 0.85 | 1 |
| WDI | 0.00 | -24.88 | -24.33 | -26.36 | -26.48 | 0.01 | 0.12 | 0.37 | 1 |
| **Congo, Rep.** | | | | | | | | | |
| PWT6 | 0.00 | -16.21 | -15.32 | -17.17 | -16.96 | 0.46 | 0.11 | 0.85 | 1 |
| PWT7 | 0.03 | -16.91 | -15.60 | -18.28 | -17.65 | 0.50 | 0.41 | 0.41 | 1 |
| PWT8 | 0.00 | -17.55 | -16.44 | -19.06 | -18.63 | 0.10 | 0.03 | 0.27 | 1 |
| WDI | 0.00 | -16.17 | -15.63 | -18.08 | -18.23 | 0.02 | 0.97 | 0.05 | 1 |
| **Madagascar** | | | | | | | | | |
| PWT6 | 0.00 | -26.10 | -24.96 | -27.74 | -27.29 | 0.19 | 0.00 | 0.01 | 1 |
| PWT7 | 0.02 | -25.00 | -23.73 | -26.64 | -26.06 | 0.72 | 0.05 | 0.18 | 1 |
| PWT8 | 0.01 | -24.67 | -23.47 | -26.32 | -25.79 | 0.83 | 0.52 | 0.07 | 1 |
| WDI | 0.02 | -24.14 | -22.86 | -25.10 | -24.50 | 0.15 | 0.02 | 0.01 | 1 |
| **Rwanda** | | | | | | | | | |
| PWT6 | 0.00 | -20.53 | -19.89 | -21.90 | -21.94 | 0.12 | 0.29 | 0.23 | 1 |
| PWT7 | 0.00 | -18.30 | -17.44 | -20.36 | -20.17 | 0.81 | 0.45 | 0.31 | 2 |
| PWT8 | 0.00 | -20.24 | -19.04 | -22.29 | -21.78 | 0.26 | 0.02 | 0.11 | 2 |
| WDI | 0.00 | -21.15 | -20.41 | -22.65 | -22.60 | 0.74 | 0.20 | 0.20 | 1 |
| **Senegal** | | | | | | | | | |
| PWT6 | 0.26 | -26.71 | -25.09 | -27.96 | -27.03 | 0.80 | 0.92 | 0.25 | 1 |
| PWT7 | 0.17 | -24.96 | -23.40 | -26.62 | -25.76 | 0.15 | 0.29 | 0.86 | 2 |
| PWT8 | 0.35 | -26.62 | -27.87 | -27.87 | -26.89 | 0.81 | 0.90 | 0.39 | 1 |
| WDI | 0.01 | -24.71 | -23.54 | -25.96 | -25.48 | 0.07 | 0.39 | 0.75 | 1 |
| **Benin** | | | | | | | | | |
| PWT6 | 0.07 | -25.78 | -24.39 | -27.26 | -26.54 | 0.36 | 0.41 | 0.71 | 1 |
| PWT7 | 0.00 | -25.02 | -23.75 | -26.50 | -25.91 | 0.34 | 0.34 | 0.14 | 1 |
| PWT8 | 0.01 | -24.58 | -23.35 | -26.06 | -25.51 | 0.85 | 0.75 | 0.37 | 1 |
| WDI | 0.00 | -24.12 | -22.88 | -25.06 | -24.50 | 0.15 | 0.03 | 0.03 | 1 |
| **Lesotho** | | | | | | | | | |
| PWT6 | 0.00 | -20.49 | -19.80 | -22.32 | -22.29 | 0.52 | 0.16 | 0.48 | 1 |
| PWT7 | 0.00 | -21.50 | -20.58 | -23.20 | -22.93 | 0.84 | 0.77 | 0.83 | 1 |
| PWT8 | 0.00 | -22.85 | -22.13 | -24.69 | -24.62 | 0.42 | 0.06 | 0.08 | 1 |
| WDI | 0.02 | -20.79 | -19.57 | -22.23 | -21.66 | 0.82 | 0.23 | 0.94 | 1 |

*Continued on next page*

Table 5 (continued)

| | | Schwarz | | Hannan-Quinn | | Autocorrelation | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Data* | *LR* | $k^*$ | $k^*+1$ | $k^*$ | $k^*+1$ | $k^*,1$ | $k^*+1,1$ | $k^*+1,2$ | $k^*$ |
| **Mauritania** | | | | | | | | | |
| PWT6 | 0.00 | -17.80 | -16.84 | -19.01 | -18.73 | 0.04 | 0.37 | 0.28 | 1 |
| PWT7 | 0.11 | -18.05 | -16.60 | -19.13 | -18.35 | 0.74 | 0.21 | 0.57 | 1 |
| PWT8 | 0.01 | -17.76 | -16.53 | -18.98 | -18.42 | 0.71 | 0.48 | 0.99 | 1 |
| WDI | 0.07 | -16.96 | -15.55 | -18.31 | -17.57 | 0.30 | 0.49 | 0.70 | 1 |
| **Togo** | | | | | | | | | |
| PWT6 | 0.00 | -21.42 | -20.47 | -22.50 | -22.23 | 0.06 | 0.47 | 0.34 | 1 |
| PWT7 | 0.00 | -21.24 | -20.53 | -22.59 | -22.55 | 0.13 | 0.91 | 0.35 | 1 |
| PWT8 | 0.00 | -21.02 | -20.33 | -22.37 | -22.35 | 0.13 | 0.80 | 0.31 | 1 |
| WDI | 0.02 | -19.21 | -17.97 | -20.56 | -19.99 | 0.08 | 0.83 | 0.43 | 1 |

Notes: The reported values are $p$-values with the exception of the Schwarz Criterion and the Hannan-Quinn Criterion. $k^*$ is the eventually inferred optimal lag length.
Source: Author's calculations.

and there is some evidence for residual autocorrelation in the first lag. In the light of otherwise reasonable properties of the model (discussed in the next section), I stick to $k = 1$ as indicated by SC in these cases, but note it as a possible caveat when interpreting the results.

**Tests for Misspecification** Before proceeding to the choice of the cointegration rank, it is worth checking the models for misspecification, as not only the trace test, but the VAR model as such, rely on a number of assumptions - most crucially, that of normal and independent residuals. Table 6 reports the p-values for tests for autocorrelation in the second lag in the final model (the first lag having been considered and reported as a criterion for the choice of the lag length before, see table 5), multivariate normality, and autoregressive conditional heteroskedasticity (ARCH) in the first and second lag. The last column reports the trace correlation, roughly interpretable as an average $R^2$ over the five equations in the model, thus summarising its overall fit.

Note that there is no indication for residual autocorrelation in the second lag in any of the models when applying a 5% significance level, and only in three cases (Gabon PWT7, Benin PWT8, Madagascar PWT6) the test would reject at the 10% level (first column, table 6). Coupled with the overall good results with respect to the first lag reported in table 5, residual autocorrelation does not appear to be a typical problem in the models.

The second column reports the $p$-values resulting for the test of multivariate normality suggested by Doornik and Hansen (2008), based on a transformation of the relevant moments (kurtosis and skewness) proposed by Shenton and Bowman (1977). In the final models, multivariate normality is almost never rejected at the 5% level. The exceptions are Kenya PWT8, and Lesotho in all datasets but PWT8, which also includes the PWT6 model, identical in its specification to JMT's and yielding output almost identical to the second decimal place (meaning that the impact of differences in the aid data is negligible, and the original specification almost certainly relied on the same test statistic). In these cases, no sensible variations in the specification could rectify the issue. This illustrates the limits of heterogeneity even in this comparatively highly flexible framework: A VAR including the present set of variables has limited validity in these cases, and more fundamental changes in the framework may be required to adequately represent the

Table 6: Misspecification tests

| *Country / Data* | *Autocorr.* | *Norm.* | ARCH(1) | ARCH(2) | *Trace Corr.* |
|---|---|---|---|---|---|
| **Burkina Faso** | | | | | |
| PWT6 | 0.35 | 0.42 | 0.05 | 0.02 | 0.40 |
| PWT7 | 0.37 | 0.11 | 0.32 | 0.07 | 0.63 |
| PWT8 | 0.42 | 0.11 | 0.21 | 0.05 | 0.62 |
| WDI | 0.72 | 0.59 | 0.10 | 0.08 | 0.51 |
| **Cameroon** | | | | | |
| PWT6 | 0.59 | 0.32 | 0.20 | 0.01 | 0.32 |
| PWT7 | 0.47 | 0.28 | 0.03 | 0.04 | 0.28 |
| PWT8 | 0.73 | 0.36 | 0.01 | 0.02 | 0.30 |
| WDI | 0.48 | 0.38 | 0.01 | 0.02 | 0.36 |
| **Gabon** | | | | | |
| PWT6 | 0.46 | 0.09 | 0.00 | 0.21 | 0.55 |
| PWT7 | 0.07 | 0.31 | 0.00 | 0.05 | 0.68 |
| PWT8 | 0.23 | 0.18 | 0.00 | 0.04 | 0.64 |
| WDI | 0.11 | 0.30 | 0.00 | 0.03 | 0.65 |
| **Kenya** | | | | | |
| PWT6 | 0.51 | 0.28 | 0.65 | 0.29 | 0.52 |
| PWT7 | 0.45 | 0.25 | 0.54 | 0.50 | 0.19 |
| PWT8 | 0.91 | 0.01 | 0.36 | 0.01 | 0.59 |
| WDI | 0.90 | 0.31 | 0.17 | 0.08 | 0.56 |
| **Congo, Rep.** | | | | | |
| PWT6 | 0.32 | 0.05 | 0.52 | 0.31 | 0.41 |
| PWT7 | 0.43 | 0.32 | 0.06 | 0.02 | 0.44 |
| PWT8 | 0.64 | 0.33 | 0.04 | 0.02 | 0.51 |
| WDI | 0.25 | 0.06 | 0.00 | 0.01 | 0.68 |
| **Madagascar** | | | | | |
| PWT6 | 0.07 | 0.23 | 0.00 | 0.01 | 0.70 |
| PWT7 | 0.34 | 0.08 | 0.01 | 0.08 | 0.65 |
| PWT8 | 0.05 | 0.15 | 0.00 | 0.00 | 0.64 |
| WDI | 0.11 | 0.42 | 0.02 | 0.02 | 0.35 |
| **Rwanda** | | | | | |
| PWT6 | 0.23 | 0.25 | 0.04 | 0.05 | 0.52 |
| PWT7 | 0.11 | 0.25 | 0.34 | 0.06 | 0.63 |
| PWT8 | 0.27 | 0.45 | 0.46 | 0.11 | 0.66 |
| WDI | 0.04 | 0.18 | 0.63 | 0.15 | 0.59 |
| **Senegal** | | | | | |
| PWT6 | 0.88 | 0.17 | 0.55 | 0.30 | 0.47 |
| PWT7 | 0.28 | 0.60 | 0.06 | 0.13 | 0.41 |
| PWT8 | 0.95 | 0.11 | 0.54 | 0.35 | 0.46 |
| WDI | 0.98 | 0.37 | 0.06 | 0.05 | 0.39 |

Continued on next page

Table 6 (continued)

| *Country / Data* | *Autocorr.* | *Norm.* | ARCH(1) | ARCH(2) | *Trace Corr.* |
|---|---|---|---|---|---|
| **Benin** | | | | | |
| PWT6 | 0.29 | 0.21 | 0.01 | 0.02 | 0.65 |
| PWT7 | 0.32 | 0.23 | 0.00 | 0.01 | 0.58 |
| PWT8 | 0.07 | 0.14 | 0.01 | 0.01 | 0.58 |
| WDI | 0.11 | 0.42 | 0.02 | 0.02 | 0.35 |
| **Lesotho** | | | | | |
| PWT6 | 0.10 | 0.00 | 0.67 | 0.03 | 0.63 |
| PWT7 | 0.57 | 0.01 | 0.53 | 0.66 | 0.62 |
| PWT8 | 0.21 | 0.06 | 0.02 | 0.05 | 0.67 |
| WDI | 0.14 | 0.00 | 0.05 | 0.03 | 0.57 |
| **Mauritania** | | | | | |
| PWT6 | 0.10 | 0.37 | 0.23 | 0.06 | 0.38 |
| PWT7 | 0.99 | 0.21 | 0.63 | 0.02 | 0.37 |
| PWT8 | 0.91 | 0.41 | 0.26 | 0.03 | 0.38 |
| WDI | 0.62 | 0.18 | 0.45 | 0.30 | 0.55 |
| **Togo** | | | | | |
| PWT6 | 0.34 | 0.55 | 0.06 | 0.05 | 0.44 |
| PWT7 | 1.00 | 0.92 | 0.21 | 0.17 | 0.53 |
| PWT8 | 0.99 | 0.91 | 0.16 | 0.17 | 0.53 |
| WDI | 0.42 | 0.06 | 0.24 | 0.06 | 0.48 |

Notes: All reported values are $p$-values, with the exception of *Trace Corr.*, which is the trace correlation. *Autocorr., Norm.*, and ARCH are tests for no residual autocorrelation, multivariate normality and no ARCH effects respectively, further discussed in the main text.

Source: Author's calculations.

macroeconomic dynamics described by the data. As the focus of the study at hand is to assess the robustness of the results within the framework provided by JMT, we refrain from more fundamental changes to the specification.

Similarly, in many models there is some evidence for the presence of ARCH effects (column 3 and 4), both in the PWT6 models consistent with JMT, and in the re-specified ones. This could be another factor undermining the performance of the trace test, which has however been shown to be relatively robust in this respect (Rahbek et al., 2002).

The trace correlation, reported in the last column, is quite persistent within each country and across datasets, indicating that the re-specified models typically reach a similar fit as the ones derived and employed by JMT.

**Cointegration rank** $r$ Perhaps the most influential and often contentious choice is that of the cointegration rank, that is, the rank $r$ of the long-run coefficient matrix $\mathbf{\Pi}$. Given the short sample period, the standard procedure for the determination, the trace test (Johansen, 1988) has very low power in the current set-up (JMT, p. 14). It therefore fails to reject unit roots at ranks that are both economically and statistically implausible, meaning that it will tend to indicate unreasonably low ranks.

In line with JMT, a number of criteria are employed in order to assure a well-grounded choice of $r$. Apart from the aforementioned (i) trace test, I will base my choice of rank

on (ii) the largest unrestricted roots of the companion matrix, (iii) the $t$-ratios of the $\alpha$-coefficients, and (iv) a visual inspection of the graphs of the cointegrating relations. These are the same criteria employed by JMT (p. 14); for a more comprehensive discussion, refer to Juselius (2006, chapter 8.5).

As reported in the first column of table 7, and in line with the above mentioned low power of the trace test for cointegration (i), it systematically suggests ranks that are equal to, or lower, than the preferred rank that I eventually determine after consideration of criteria (ii)-(iv). The only exceptions to this occur for Lesotho PWT6, PWT7 and PWT8, as well as Madagascar in PWT6. This is not surprising in the light of the results discussed in the previous section, where the residuals in these models have been found likely to violate the assumption of normality, on which the test fundamentally relies.[12] I emphasise once more that this is pervasive throughout both the re-specified models, as well as those employed by JMT.[13]

The (ii) largest unrestricted roots are reported in columns 4 and 5 of table 7, for both the eventually preferred rank $r^*$, and for $r^* + 1$. The idea here is that the largest unrestricted root in the system should be significantly smaller than one, in order to ensure a stationary process $\Delta \mathbf{X}_t$. Practically speaking, the largest unrestricted root at $r^*$ should be small, while the one at $r^* + 1$ should be close to one, as otherwise the rank could be confidently increased, preserving the information concerning another equilibrium relationship. It is, however, only indicative again, especially because the confidence intervals of the roots are unknown, and there is therefore no hard criterion in order to determine whether a root is significantly smaller than one (Juselius, 2006, p. 143). Nevertheless, the largest roots at $r^* + 1$ are in the vast majority of models substantially larger than those at $r^*$, providing some justification for this choice. Most importantly in the context of a robustness check, they tend to be of a similar amplitude as those obtained in the PWT6 (JMT) models. For illustration, appendix D plots the roots of the companion matrix for all ranks for Burkina Faso in PWT6.

The (iii) largest $t$-ratio in the $\alpha$-vector associated with the $r^*$'th CI relation (column 5) gives an indication about the relevance of the last potential equilibrium relationship included in the model. The same figure for the $r^*+1$'th $\alpha$-vector (column 6) provides such an indication for the first vector dismissed from the analysis. Juselius (2006) proposes a threshold of about $|t| > 2.6$, which is surpassed in most specifications, as $r^*+1$ is typically found to be susceptible of non-stationarity. Where this is not the case, $r^* + 1$ is generally determined as the second best choice of rank, reported in parentheses in the last column.

In practice, the (iv) visual inspection of the graphs of the CI relationships turns out to be the most common tie-breaker. While in principle this is quite a subjective criterion, in most cases there tends to be a rather sharp difference in the appearance of the graphs of the $r^*$'th CI (the CIs being ordered by the corresponding eigenvalues) and the $r^*+1$'th (or, if this is the second best choice, the $r^* + 2$'th). The highest included CI typically resembles a white noise process, while larger order CIs have a distinctly persistent, that is, non-stationary appearance. The values reported in the second last column of table 7 report the rank that can be best justified based on the graphs, followed by the best alternative rank choice resulting from them in parentheses; In cases where the difference is quite clear-cut, the alternative rank will be below the preferred choice. Where the

[12]The table does not report a $p$-value in the cases where the trace test concludes full rank $r = p = 5$, as this inference emerges from the rejection of all lower ranks versus the alternative hypothesis of $r = p$.

[13]The relevant output underlying the original choices of rank, kindly provided by the authors, confirms this.

Table 7: Criteria for the choice of rank

| | Trace test | | Largest root | | max $\|t\|$ in $\boldsymbol{\alpha}_{r^*}$ | | | |
|---|---|---|---|---|---|---|---|---|
| *Country / Data* | *Inf.* | *p-value* | $r^*$ | $r^*+1$ | $r^*$ | $r^*+1$ | *Graph* | $r^*(r')$ |
| **Burkina Faso** | | | | | | | | |
| PWT6 | 0 | 0.52 | 0.32 | 0.86 | 2.98 | 3.12 | 1(2) | 2(1) |
| PWT7 | 2 | 0.21 | 0.59 | 0.63 | -4.1 | -4.54 | 2(1) | 2(1) |
| PWT8 | 2 | 0.17 | 0.54 | 0.61 | -4.17 | 4.42 | 2(1) | 2(1) |
| WDI | 1 | 0.18 | 0.34 | 0.63 | 4.51 | 3.21 | 2(3) | 3(2) |
| **Cameroon** | | | | | | | | |
| PWT6 | 0 | 0.08 | 0.63 | 0.74 | -4.66 | 1.34 | 3(2) | 3(2) |
| PWT7 | 0 | 0.55 | 0.64 | 0.68 | -2.44 | -1.18 | 2(3) | 3(2) |
| PWT8 | 0 | 0.25 | 0.66 | 0.69 | 2.55 | 1.44 | 2(3) | 3(2) |
| WDI | 1 | 0.11 | 0.6 | 0.73 | 3.84 | -2.59 | 3(2) | 3(2) |
| **Gabon** | | | | | | | | |
| PWT6 | 1 | 0.21 | 0.92 | 0.92 | 4.16 | -2.89 | 3(4) | 3(4) |
| PWT7 | 3 | 0.19 | 0.38 | 0.44 | 7.66 | -3.46 | 3(4) | 3(4) |
| PWT8 | 2 | 0.16 | 0.87 | 0.99 | -3.67 | -4.57 | 3(4) | 3(2) |
| WDI | 2 | 0.08 | 0.93 | 0.99 | 5.92 | -2.98 | 3(4) | 3(2) |
| **Kenya** | | | | | | | | |
| PWT6 | 1 | 0.20 | 0.61 | 0.83 | 4.63 | 4.31 | 2(3) | 3(2) |
| PWT7 | 1 | 0.06 | 0.67 | 0.85 | -2.86 | -3.71 | 2(3) | 3(2) |
| PWT8 | 3 | 0.22 | 0.55 | 0.85 | 3.53 | -4.49 | 2(3) | 3(2) |
| WDI | 3 | 0.08 | 0.64 | 0.78 | 4.41 | 5.02 | 4(3) | 3(4) |
| **Congo, Rep.** | | | | | | | | |
| PWT6 | 1 | 0.52 | 0.64 | 0.77 | 3.65 | 3.71 | 2(3) | 2(3) |
| PWT7 | 1 | 0.43 | 0.66 | 0.81 | 2.83 | -3.83 | 3(2) | 3(2) |
| PWT8 | 1 | 0.26 | 0.69 | 0.80 | 3.18 | 3.75 | 2(3) | 3(2) |
| WDI | 4 | 0.28 | 0.53 | 0.91 | 3.58 | -2.54 | 3(4) | 4(3) |
| **Madagascar** | | | | | | | | |
| PWT6 | 5 | 0.00 | 0.52 | 0.63 | -2.71 | -2.96 | 4(3) | 4(3) |
| PWT7 | 3 | 0.31 | 0.51 | 0.64 | 2.87 | -2.50 | 3(4) | 4(3) |
| PWT8 | 3 | 0.38 | 0.21 | 0.66 | -6.58 | -3.83 | 3(2) | 3(2) |
| WDI | 1 | 0.14 | 0.35 | 0.79 | -3.61 | -2.00 | 2(3) | 2(3) |
| **Rwanda** | | | | | | | | |
| PWT6 | 2 | 0.27 | 0.48 | 0.50 | -3.64 | 3.13 | 2(3) | 3(2) |
| PWT7 | 1 | 0.72 | 0.55 | 0.89 | -3.2 | -2.52 | 2(1) | 2(1) |
| PWT8 | 1 | 0.35 | 0.58 | 0.87 | 5.73 | -2.67 | 2(1) | 2(1) |
| WDI | 2 | 0.24 | 0.67 | 0.70 | -4.23 | 3.02 | 3(2) | 3(2) |
| **Senegal** | | | | | | | | |
| PWT6 | 2 | 0.14 | 0.28 | 0.86 | -3.51 | 2.9 | 2(3) | 3(2) |
| PWT7 | 0 | 0.51 | 0.21 | 0.87 | 3.45 | 2.22 | 3(2) | 3(2) |
| PWT8 | 2 | 0.17 | 0.32 | 0.84 | 3.28 | 2.84 | 2(3) | 3(2) |
| WDI | 1 | 0.18 | 0.59 | 0.8 | -2.73 | -2.27 | 2(3) | 2(3) |
| **Benin** | | | | | | | | |
| PWT6 | 3 | 0.23 | 0.36 | 0.65 | -4.83 | 3.18 | 3(4) | 3(4) |
| PWT7 | 2 | 0.33 | 0.31 | 0.82 | 4.23 | -1.45 | 3(2) | 3(2) |
| PWT8 | 2 | 0.10 | 0.24 | 0.63 | 3.97 | -3.12 | 3(2) | 3(2) |
| WDI | 0 | 0.16 | 0.25 | 0.35 | -4.34 | -3.61 | 2(1) | 2(1) |

Continued on next page

Table 7 (continued)

| | Trace test | | Largest root | | max $\|t\|$ in $\boldsymbol{\alpha}_{r^*}$ | | | |
|---|---|---|---|---|---|---|---|---|
| *Country / Data* | *Inf.* | *p-value* | $r^*$ | $r^*+1$ | $r^*$ | $r^*+1$ | *Graph* | $r^*(r')$ |
| **Lesotho** | | | | | | | | |
| PWT6 | 5 | 0.00 | 0.85 | 0.89 | 4.6 | 4.46 | 1(2) | 3(2) |
| PWT7 | 4 | 0.07 | 0.55 | 0.83 | 5.02 | -3.53 | 2(1) | 2(1) |
| PWT8 | 5 | 0.00 | 0.55 | 0.6 | 5.76 | 5.01 | 2(3) | 3(2) |
| WDI | 2 | 0.37 | 0.58 | 0.82 | 5.34 | -3.15 | 1(2) | 2(1) |
| **Mauritania** | | | | | | | | |
| PWT6 | 0 | 0.23 | 0.53 | 0.63 | -3.17 | -2.68 | 2(3) | 3(2) |
| PWT7 | 0 | 0.20 | 0.41 | 0.77 | -4.54 | -2.94 | 3(2) | 3(2) |
| PWT8 | 0 | 0.33 | 0.45 | 0.62 | -4.27 | -4.08 | 2(3) | 2(3) |
| WDI | 2 | 0.41 | 0.00 | 0.61 | 4.09 | 3.28 | 2(1) | 2(1) |
| **Togo** | | | | | | | | |
| PWT6 | 2 | 0.41 | 0.34 | 0.68 | -4.96 | -3.01 | 2(1) | 2(3) |
| PWT7 | 3 | 0.26 | 0.59 | 0.66 | 4.53 | -1.85 | 3(2) | 3(2) |
| PWT8 | 3 | 0.17 | 0.63 | 0.68 | -4.93 | 1.95 | 3(2) | 3(2) |
| WDI | 2 | 0.62 | 0.54 | 0.78 | 3.43 | 1.62 | 3(2) | 3(2) |

Notes: *Trace test* reports the rank suggested by Johansen (1988)'s trace test with the corresponding $p$-value of acceptance. *Largest root* and *max* $|t|$ *in* $\boldsymbol{\alpha}_{r^*}$ report the respective values for the preferred rank and the one above, *Graph* indicates the rank most confidently suggested by the graph and the best alternative in parentheses, $r^*(r')$ the inferred preferred and second best choice of rank.
Source: Author's calculations.

next-highest CI also appears to be acceptable, e.g., has a short period of persistence but otherwise looks stationary, it may be reported as the suggested second best rank. Appendix E plots the cointegrating relations for all ranks for Burkina Faso in PWT6.

The last column reports the final choice of ranks after weighting of criteria (i)-(iv).[14] It is apparent from the discussion in this section that the choice of the cointegration rank is everything but straightforward, and the researcher faces significant trade-offs in the process. This provides justification for JMT's procedure of assessing the results from two different economic angles, establishing transparency by essentially picking the rank that yields the results most consistent with the respective prior of effectiveness or harmfulness, and reporting both.

Table 8 reports the final model choices, stating their respective lag-length, dummy variables, and first and second best choices of cointegration ranks.

### 5.3.2 Results of the re-specified models

The results emerging from these specifications are summarised in table 9, the underlying $t$-ratios are reported in appendix F. As in table 3, $+$ and $-$ stand for positive and negative coefficients, while the subscripts 0 and 00 denote coefficients with absolute $t$-ratios below 2.0 and 1.6 respectively. The table is organised in two columns for each dataset, reporting

[14]The choice of 3(2) for Togo PWT6, a model repeatedly found to be problematic in the previous sections, may deserve some discussion as it seems at odds with the indications provided by the criteria. The main rationale behind the choice of rank is to preserve consistency with JMT, even though their choice may appear rather surprising in this particular instance. As noted earlier, JMT's output is very much in line with the results I obtain; this can be verified for the trace test and the roots of the companion matrix.

Table 8: Final model specifications

| *Country* | *Lags* | *Dummy variables* | $r^*$ | $r'$ |
|---|---|---|---|---|
| **Stable countries** | | | | |
| **Burkina Faso** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | $Dtr71_t, Dp00_t$ | 2 | 1 |
| PWT7 | 1 | $Dtr71_t, Ds74_t, Dp79_t, Dp85_t, Dp94_t, Dp00_t$ | 2 | 1 |
| PWT8 | 1 | $Dtr71_t, Ds74_t, Dp79_t, Dp85_t, Dp94_t, Dp00_t$ | 2 | 1 |
| WDI | 1 | $Dp85_t, Dp94_t, Dp00_t$ | 2 | 3 |
| **Cameroon** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | None | 3 | 2 |
| PWT7 | 1 | None | 3 | 2 |
| PWT8 | 1 | None | 3 | 2 |
| WDI | 1 | None | 3 | 2 |
| **Gabon** | *Sample: 1965-2002* | | | |
| PWT6 | 1 | $Dtr7478_t, Dtr8900_t$ | 3 | 4 |
| PWT7 | 1 | $Dtr7478_t, Dp87_t, Dp00_t$ | 3 | 2 |
| PWT8 | 1 | $Dtr7478_t, Dp87_t, Dtr0002_t$ | 3 | 2 |
| WDI | 1 | $Dtr7478_t, Dp87_t, Dtr0002_t$ | 3 | 2 |
| **Kenya** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | $Ds68_t, Ds94_t$ | 3 | 2 |
| PWT7 | 1 | $Ds68_t, Ds94_t$ | 3 | 2 |
| PWT8 | 1 | $Ds68_t, Dtr9900_t, Ds94_t$ | 3 | 2 |
| WDI | 1 | $Ds70_t, Ds94_t$ | 3 | 4 |
| **Intermediate countries** | | | | |
| **Congo, Rep.** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | $Dp94_t, Dp05_t$ | 2 | 3 |
| PWT7 | 1 | $Ds96_t, Dp05_t$ | 3 | 2 |
| PWT8 | 1 | $Dp91_t, Ds96_t, Dp05_t$ | 3 | 2 |
| WDI | 1 | $Dp77_t, Dp94_t, Ds96_t, Dp98_t, Dp00_t, Dp05_t$ | 4 | 3 |
| **Madagascar** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | $Ds75_t, Ds82_t, Dp94_t$ | 4 | 3 |
| PWT7 | 1 | $Ds82_t, Dtr85_t, Ds94_t$ | 4 | 3 |
| PWT8 | 1 | $Ds82_t, Dp89_t, Ds94_t$ | 3 | 2 |
| WDI | 1 | None | 2 | 3 |
| **Rwanda** | *Sample: 1960-2007* | | | |
| PWT6 | 1 | $Ds94_t, Dp81_t$ | 3 | 2 |
| PWT7 | 2 | $Ds94_t, Dp63_t$ | 2 | 1 |
| PWT8 | 2 | $Ds94_t, Dp63_t$ | 2 | 1 |
| WDI | 1 | $Ds94_t, Dtr6375_t.Dp97_t$ | 3 | 2 |
| **Senegal** | *Sample: 1960-2008* | | | |
| PWT6 | 1 | $Ds69_t$ | 3 | 2 |
| PWT7 | 2 | none | 3 | 2 |
| PWT8 | 1 | $Ds69_t$ | 3 | 2 |
| WDI | 1 | $Ds69_t$ | 2 | 3 |

*Continued on next page*

Table 8 (continued)

| *Country* | *Lags* | *Dummy variables* | $r^*$ | $r'$ |
|---|---|---|---|---|
| **Unstable countries** | | | | |
| **Benin** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | $Dp75_t, Ds82_t, Dp94_t$ | 3 | 4 |
| PWT7 | 1 | $Dp82_t, Ds86_t, Dp94_t$ | 3 | 2 |
| PWT8 | 1 | $Ds82_t, Dp89_t, Dp94_t$ | 3 | 2 |
| WDI | 1 | None | 2 | 1 |
| **Lesotho** | *Sample: 1963-2007* | | | |
| PWT6 | 1 | $Dtr65_t, Ds70_t, Ds78_t, Dp83_t, Dp99_t$ | 3 | 2 |
| PWT7 | 1 | $Dtr65_t, Ds71_t, Dp78_t, Dp83_t, Dp99_t$ | 2 | 1 |
| PWT8 | 1 | $Ds71_t, Dp75_t, Ds78_t, Dp83_t, Dp99_t$ | 3 | 2 |
| WDI | 1 | $Ds71_t, Dp75_t, Ds78_t, Dp83_t, Dp98_t, Dp99_t$ | 2 | 1 |
| **Mauritania** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | $Ds92_t$ | 3 | 2 |
| PWT7 | 1 | $Dp86_t$ | 3 | 2 |
| PWT8 | 1 | $Dp86_t, Dp95_t$ | 2 | 3 |
| WDI | 1 | $Dp76_t, Dp99_t, Dp03_t$ | 2 | 1 |
| **Togo** | *Sample: 1965-2007* | | | |
| PWT6 | 1 | $Dp93_t$ | 2 | 3 |
| PWT7 | 1 | $Dp86_t, Dp87_t, Dp93_t$ | 3 | 2 |
| PWT8 | 1 | $Dp86_t, Dp87_t, Dp93_t$ | 3 | 2 |
| WDI | 1 | $Dp71_t, Dp75_t, Dp93_t$ | 3 | 2 |

Notes: The table summarises all models derived in the re-specification exercise. $Dp, Dtr, Ds$ are permanent, transitory and shift dummies respectively, as defined in the main text. $r^*$ and $r'$ are the first and second best choices of rank respectively.

the results under the prior of effectiveness and the prior of harmfulness respectively, and four rows per country, one for each of the macrovariables under consideration. The first, second and third panel comprise the four countries where the results were most stable, relatively stable and least stable in the first replication exercise (see table 8).

The last two columns of table 9 count the number of coefficients that are consistent with the ones obtained from PWT6[15], where 'consistency' is still defined as in table 3. An apparent pattern is that the countries that yielded particularly consistent results under the original JMT models also remain substantially more consistent under the re-specified models, with 43 out of 48 consistent coefficients under the prior of aid effectiveness. This compares to 16 stable coefficients in both the countries with the least stable results previously, and the intermediate ones. Overall, 87 out of 144 possible coefficients (or 60%) of the coefficients remained stable. This pattern is repeated under the prior of harmfulness in a slightly less pronounced manner, and at a generally lower level of consistency. The datasets agree for 22 coefficients in the stable countries, 26 in the intermediate, and for 12 in the least stable ones. Overall, this sums up to only 60 out of 144 coefficients (42%) that remain stable under the prior of harmfulness.

[15]Note that while these are overall consistent with the results obtained by JMT, but may differ in some cases; where this is the case, I use my own results as a benchmark for the sake of internal consistency. This affects Burkina Faso, the Republic of Congo and Madagascar, where differences in the aid data also lead me to a different specification than JMT.

Table 9: Results of the re-specified models

| | | PWT6 | | PWT7 | | PWT8 | | WDI | | Stable | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Eff. | Harm. | Eff. | Harm. | Eff. | Harm. | Eff. | Harm. | Eff. | Harm. |
| Burkina Faso | $y_t$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | 3 | 2 |
| | $inv_t$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $-$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $-$ | 2 | 1 |
| | $c_t$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $+_{0}$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $+$ | 3 | 1 |
| | $g_t$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $+_{0}$ | $+_{00}$ | $+_{00}$ | $+_{0}$ | $-$ | 2 | 1 |
| Cameroon | $y_t$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-$ | 3 | 2 |
| | $inv_t$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | 3 | 3 |
| | $c_t$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-$ | 3 | 2 |
| | $g_t$ | $-_{00}$ | $-$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | 3 | 0 |
| Gabon | $y_t$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $+_{0}$ | $+_{0}$ | $+_{00}$ | $-_{00}$ | 2 | 2 |
| | $inv_t$ | $+_{00}$ | $-_{00}$ | $+_{00}$ | $-_{00}$ | $+_{0}$ | $+_{0}$ | $+_{00}$ | $-_{00}$ | 2 | 2 |
| | $c_t$ | $+_{00}$ | $-$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | 3 | 0 |
| | $g_t$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | 3 | 3 |
| Kenya | $y_t$ | $+$ | $+_{00}$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | 3 | 0 |
| | $inv_t$ | $+$ | $+$ | $+$ | $+_{0}$ | $+$ | $+_{0}$ | $+$ | $+$ | 3 | 1 |
| | $c_t$ | $+$ | $+_{0}$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | 3 | 0 |
| | $g_t$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $+_{0}$ | $-$ | 2 | 2 |
| Congo, Rep. | $y_t$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | 3 | 3 |
| | $inv_t$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | 3 | 3 |
| | $c_t$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | 3 | 3 |
| | $g_t$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | 3 | 3 |
| Mada-gascar | $y_t$ | $+$ | $+_{00}$ | $-_{00}$ | $+_{00}$ | $+_{0}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | 1 | 3 |
| | $inv_t$ | $+$ | $-_{00}$ | $-_{00}$ | $-$ | $+$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | 1 | 2 |
| | $c_t$ | $+$ | $+$ | $+_{00}$ | $+$ | $+$ | $+$ | $-_{00}$ | $-_{00}$ | 1 | 2 |
| | $g_t$ | $-$ | $-$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | 0 | 0 |
| Rwanda | $y_t$ | $+_{00}$ | $-$ | $+$ | $+_{0}$ | $+$ | $+$ | $+$ | $+$ | 0 | 0 |
| | $inv_t$ | $+$ | $-$ | $+$ | $+_{0}$ | $+_{0}$ | $+_{0}$ | $+$ | $+$ | 2 | 0 |
| | $c_t$ | $+$ | $+_{00}$ | $+_{00}$ | $+$ | $+$ | $+$ | $+$ | $+$ | 2 | 0 |
| | $g_t$ | $-_{0}$ | $-_{00}$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $+_{00}$ | 0 | 3 |
| Senegal | $y_t$ | $+$ | $+$ | $+$ | $+_{00}$ | $+$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | 2 | 0 |
| | $inv_t$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | 2 | 3 |
| | $c_t$ | $+$ | $+$ | $+$ | $+_{0}$ | $+$ | $+_{0}$ | $+$ | $+$ | 3 | 1 |
| | $g_t$ | $+$ | $+$ | $+$ | $+_{0}$ | $+$ | $+_{0}$ | $+_{00}$ | $+_{00}$ | 2 | 0 |
| Benin | $y_t$ | $+_{00}$ | $-$ | $+$ | $-_{00}$ | $+$ | $-_{00}$ | $-_{00}$ | $-_{0}$ | 1 | 0 |
| | $inv_t$ | $+$ | $-$ | $+$ | $-$ | $+$ | $+_{00}$ | $+_{00}$ | $-$ | 2 | 2 |
| | $c_t$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | $-_{00}$ | $+_{00}$ | 2 | 2 |
| | $g_t$ | $-$ | $-$ | $+$ | $-_{00}$ | $-$ | $-$ | $+_{00}$ | $+$ | 1 | 1 |
| Lesotho | $y_t$ | $+$ | $+$ | $-$ | $-$ | $+_{00}$ | $-_{00}$ | $+_{00}$ | $-$ | 0 | 0 |
| | $inv_t$ | $+$ | $+$ | $+_{00}$ | $+_{00}$ | $+$ | $+_{00}$ | $+_{00}$ | $-$ | 1 | 0 |
| | $c_t$ | $+$ | $+$ | $-$ | $-$ | $-$ | $-$ | $+_{00}$ | $-$ | 0 | 0 |
| | $g_t$ | $+$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $+_{00}$ | $-$ | $-_{0}$ | $-_{00}$ | 0 | 2 |
| Mauri-tania | $y_t$ | $-_{00}$ | $+_{00}$ | $+_{0}$ | $-$ | $-_{00}$ | $-$ | $-$ | $-$ | 1 | 0 |
| | $inv_t$ | $+$ | $-_{00}$ | $+_{00}$ | $-$ | $+_{00}$ | $-_{00}$ | $-_{00}$ | $-_{00}$ | 0 | 2 |
| | $c_t$ | $+_{00}$ | $+_{00}$ | $+$ | $-_{00}$ | $+$ | $+$ | $+_{00}$ | $+_{00}$ | 1 | 2 |
| | $g_t$ | $+$ | $+_{00}$ | $+_{00}$ | $+$ | $+_{00}$ | $+_{00}$ | $+$ | $+$ | 1 | 1 |

Table 9 (continued)

| | | PWT6 | | PWT7 | | PWT8 | | WDI | | Stable | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Eff. | Harm. | Eff. | Harm. | Eff. | Harm. | Eff. | Harm. | Eff. | Harm. |
| Togo | $y_t$ | $+$ | $+_0$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | 3 | 0 |
| | $inv_t$ | $-_{00}$ | $+_0$ | $-$ | $-$ | $-$ | $-$ | $+$ | $+_{00}$ | 0 | 0 |
| | $c_t$ | $+$ | $-_{00}$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | 3 | 0 |
| | $g_t$ | $+$ | $+$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | $+_{00}$ | 0 | 0 |
| $\Sigma$ | $y_t$ | 5 | 2 | 5 | 2 | 5 | 1 | 3 | 3 | | |
| | $inv_t$ | 6 | 2 | 3 | 5 | 5 | 1 | 3 | 3 | | |
| | $\Sigma$ | 8 | 2 | 5 | 6 | 7 | 2 | 3 | 5 | | |

Notes: $+$ indicates a positive coefficient, $-$ a negative one. The subscript '0' indicates an absolute $t$-ratio between 1.6 and 2, '00' indicates one lower than 1.6. The absence of a subscript indicates an absolute $t$-ratio $> 2$. The last three rows count the number of cases of effectiveness and harmfulness on $y_t$, $inv_t$, and overall. The last two columns count the cases of consistency under each prior.

One possible driver of the difference between the stability of the results is that the original results seem to be systematically different between stable and unstable countries: most of the coefficients of the stable countries are in fact insignificant. As noted in the previous replication exercise, insignificant coefficients had a stronger tendency to remain stable than significant ones. Note however that even Kenya, where 3 out of 4 coefficients are significant with PWT6, is almost perfectly consistent under the prior of effectiveness, and far better than any of the inconsistent countries.

The overall *conclusions* suggested by these results are summarised in the last three rows of table 9 for each of the four datasets, across the 12 countries in the sub-sample. The first two of these rows report the number of positive (negative) significant coefficients for GDP ($y_t$) and investment ($inv_t$) respectively. The last row ($\Sigma$) counts the number of countries where at least one of the two is significant and positive (negative), and where therefore aid is considered to have been effective (harmful) in the long run, overall. In the sub-sample at hand, the overall conclusion changes twice once we account for the impact of the data on the modelling process, and remains constant only once, but with less strong support. The conclusion of aid effectiveness, clearly supported by PWT6 with 8 out of 12 countries providing evidence for it, compared to only 1 country (Benin) providing evidence for harmfulness, also finds support in PWT8, where 7 countries indicate effectiveness, and 2 harmfulness. PWT7 and WDI now lend some support to the hypothesis of harmfulness: In PWT7, there is evidence of effectiveness in 5 countries, and evidence of harmfulness in 6. In WDI, these figures are 3 and 5, respectively.

Adjusting the models to the new data did therefore not help to re-establish consistency where the results turned out to be shaky when applying JMTs original models to different datasets. Instead, it seems like the divergence of the results has been even exacerbated by letting the new data shape the statistical models.

# 6 Conclusions

In this study, I used the econometric framework provided by Juselius, Møller and Tarp (2014) in order to explore the stability of macroeconomic inference drawn from methods of time series analysis with respect to inconsistencies across datasets.

This included, in a first exercise, the application of the statistical models as specified by

JMT to data from the Penn World Table versions 6.3, 7.1, 8.0, and the World Development Indicators 2015. About one third of the relevant coefficients changed qualitatively in each of the datasets applied, that is, in a way that affects inference (change of sign or significance level). Similarly, the country-wise conclusion regarding whether aid was effective, harmful or neutral, remained stable in about two thirds of the country-dataset combinations.

The second exercise allowed for a more fundamental impact of the data as I re-specified the country-specific models for the subset of 12 countries for which sufficient data from all considered sources were available. The results show that the modelling process can be influenced in a significant manner by the inconsistencies between the datasets, especially when it comes to the detection of outliers and the resulting dummy variables, and to the choice of the rank of cointegration. As a general pattern, the countries that proved to be particularly consistent in the first exercise also yielded more consistent models in the second exercise, and consequently more similar results. In the countries that were particularly inconsistent in the first exercise, the alterations to the models frequently restored significance in the results where it had been lost. These results did not necessarily correspond to the original results obtained by JMT, and overall, the re-specification exacerbated the divergence.

My results suggest that the choice of dataset can have substantial impact on the results obtained from time series analysis, and future research needs to address this. One obvious recommendation is that robustness checks with alternative data should be standard in the literature, wherever feasible. This is not always straightforward to do, especially in the context of time series analysis, as the specification may have to be fundamentally rethought when the data changes. At the same time, this only increases the potential ramifications of changes in the data and renders the exercise even more urgent. My findings do, however, suggest that it may be a useful and practical heuristic to confront one's final model specification to alternative datasets, as the results that remained robust in this exercise were also much more likely to persist after the entire model had been re-specified.

Beyond routine robustness checks, future research should also seek to establish whether alternative estimation techniques are more robust to data revisions and measurement error than the Cointegrated VAR framework employed in the study at hand. Candidates could include likelihood analysis based on error distributions with broader tails or bootstrap based inference.

It is worth noting that for most of the countries included in the analysis (one half to two thirds, depending on the criterion), the results remain unaffected by any of the differences across datasets, even though those differences often appear substantial. In the light of the recent debate on 'Poor Numbers' (Jerven, 2013c), the 'Anarchy of Numbers' (Roodman, 2007), and 'Africa's Statistical Tragedy' (Devarajan, 2013), this result indeed seems more optimistic than the tone of the debate suggests. Categorically dismissing the only available evidence on the past economic performances of African countries, as advocated by some, would represent a considerable waste of information; it is what the best estimates tell us, and in most cases the reported sums seem to be sufficiently in agreement in order to provide valuable insights about the mechanisms at work.

# References


Addison, T. and Baliamoune-Lutz, M. (2017). Aid, the real exchange rate and why policy matters: The cases of Morocco and Tunisia. *The Journal of Development Studies*, 53(7):1104–1121.

Atherton, P., Appleton, S., and Bleaney, M. (2011). Growth regressions and data revisions in Penn World Tables. *Journal of Economic Studies*, 38(3):301–312.

Breton, T. R. (2012). Penn World Table 7.0: Are the data flawed? *Economics Letters*, 117(1):208–210.

Breton, T. R. (2015). Changes in the effect of capital and TFP on output in Penn World Tables 7 and 8: Improvement or error? *Center for Research in Economics and Finance (CIEF), Working Papers, https://repository.eafit.edu.co/handle/10784/388*, (15-03).

Breton, T. R. and García, J. J. (2016). ICP 2005 construction prices: Are they underestimated in developing countries? *Review of Income and Wealth*, (62):380–393.

Bwire, T., Lloyd, T., and Morrissey, O. (2017). Fiscal reforms and the fiscal effects of aid in Uganda. *The Journal of Development Studies*, 53(7):1019–1036.

Ciccone, A. and Jarociński, M. (2010). Determinants of economic growth: will data tell? *American Economic Journal: Macroeconomics*, 2(4):222–246.

Deaton, A. and Heston, A. (2010). Understanding PPPs and PPP-based national accounts. *American Economic Journal: Macroeconomics*, 2(4):1–35.

Deaton, A. and Tten, B. (2017). Trying to understand the PPPs in ICP 2011: Why are the results so different? *American Economic Journal: Macroeconomics*, 9(1):243–264.

Devarajan, S. (2013). Africa's statistical tragedy. *Review of Income and Wealth*, 59(S1):S9–S15.

Doornik, J. A. (2009). Autometrics. In Castle, J. and Shepard, N., editors, *The Methodology and Practice of Econometrics*, chapter 4, pages 88–121. Oxford University Press.

Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70(1):927–939.

Easterly, W., Levine, R., and Roodman, D. (2004). Aid, policies, and growth: Comment. *American Economic Review*, 94(3):774–780.

Eberhardt, M. and Teal, F. (2011). Econometrics for grumblers: a new look at the literature on cross-country growth empirics. *Journal of Economic Surveys*, 25(1):109–155.

Feenstra, R. C., Inklaar, R., and Timmer, M. (2013a). Comparing PWT 8.0 with pwt 7.1. *University of California, Davis and University of Groningen.*

Feenstra, R. C., Inklaar, R., and Timmer, M. (2013b). Penn World Table v. 8.0. *University of California, Davis and University of Groningen, http://www.rug.nl/research/ggdc/data/penn-world-table.*

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review*, 105(10):3150–3182.

Gebregziabher, F. (2014). The long-run macroeconomic effects of aid and disaggregated aid in Ethiopia. *Journal of International Development*, 26(4):520–540.

Guha-Sapir, D., Below, R., and Hoyois, P. (2014). EM-DAT: International disaster database – `www.emdat.be`. *Université Catholique de Louvain, Brussels: Belgium.*

Hanushek, E. A. and Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American Economic Review*, 90(5):1184–1208.

Hausmann, R., Pritchett, L., and Rodrik, D. (2005). Growth accelerations. *Journal of Economic Growth*, 10(4):303–329.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028.

Heston, A., Summers, R., and Aten, B. (2009). Penn World Table v. 6.3. *Center for International Comparisons of Production, Income and Prices (Philadelphia: University of Pennsylvania).*

Heston, A., Summers, R., and Aten, B. (2012). Penn World Table version 7.1. *Center for International Comparisons of Production, Income, and Prices at the University of Pennsylvania.*

Hoover, K. D., Johansen, S., and Juselius, K. (2008). Allowing the data to speak freely: The macroeconometrics of the cointegrated vector autoregression. *American Economic Review*, 98(2):251–255.

Inklaar, R. and Prasada Rao, D. (2017). Cross-country income levels over time: did the developing world suddenly become much richer? *American Economic Journal: Macroeconomics*, 9(1):265–290.

Jerven, M. (2011). Users and producers of African income: Measuring the progress of African economies. *African Affairs*, 110(439):169–190.

Jerven, M. (2013a). Comparability of GDP estimates in sub-Saharan Africa: The effect of revisions in sources and methods since structural adjustment. *Review of Income and Wealth*, 59(S1):S16–S36.

Jerven, M. (2013b). For richer, for poorer: GDP revisions and Africa's statistical tragedy. *African Affairs*, 112(446):138–147.

Jerven, M. (2013c). *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It.* Cornell University Press.

Jerven, M. (2016). Discrepancies: Why do GDP growth rates differ? *Review of Agrarian Studies*, 6(1):63–80.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231–254.

Johnson, S., Larson, W., Papageorgiou, C., and Subramanian, A. (2013). Is newer better? Penn World Table revisions and their impact on growth estimates. *Journal of Monetary Economics*, 60(2):255–274.

Jones, B. F. and Olken, B. A. (2005). Do leaders matter? National leadership and growth since World War II. *The Quarterly Journal of Economics*, 120(3):835–864.

Juselius, K. (2006). *The cointegrated VAR model: methodology and applications*. Oxford University Press.

Juselius, K., Møller, N. F., and Tarp, F. (2014). The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis. *Oxford Bulletin of Economics and Statistics*, 76(2):153–184.

Juselius, K., Reshid, A., and Tarp, F. (2017). The real exchange rate, foreign aid and macroeconomic transmission mechanisms in Tanzania and Ghana. *The Journal of Development Studies*, 53(7):1075–1103.

Mascagni, G. and Timmis, E. (2017). The fiscal effects of aid in Ethiopia: Evidence from CVAR applications. *The Journal of Development Studies*, 53(7):1037–1056.

Ndulu, B. J., O'Connell, S. A., Bates, R. H., Collier, P., and Soludo, C. C., editors (2008). *The political economy of economic growth in Africa, 1960-2000*, volume 2. Cambridge University Press.

Osei, R., Morrissey, O., and Lloyd, T. (2005). The fiscal effects of aid in Ghana. *Journal of International Development*, 17(8):1037–1053.

Pettersson, T. and Wallensteen, P. (2015). Armed conflicts, 1946–2014. *Journal of Peace Research*, 52(4):536–550.

Pinkovskiy, M. and Sala-i-Martin, X. (2016). Newer need not be better: Evaluating the Penn World Tables and the World Development Indicators using nighttime lights. *NBER Working Paper*, (22216).

Ponomareva, N. and Katayama, H. (2010). Does the version of the Penn World Tables matter? An analysis of the relationship between growth and volatility. *Canadian Journal of Economics/Revue canadienne d'économique*, 43(1):152–179.

Rahbek, A., Hansen, E., and Dennis, J. G. (2002). ARCH innovations and their impact on cointegration rank testing. *Centre for Analytical Finance Working Paper, University of Copenhagen*, (22:15).

Ram, R. and Ural, S. (2013). Comparison of GDP per capita data in Penn World Table and World Development Indicators. *Social Indicators Research*, pages 1–8.

Ramey, G. and Ramey, V. (1995). Cross-country evidence on the link between volatility and growth. *American Economic Review*, 85(5):1138–1151.

Roodman, D. (2007). The anarchy of numbers: aid, development, and cross-country empirics. *The World Bank Economic Review*, 21(2):255–277.

Sala-i-Martin, X., Doppelhofer, G., and Miller, R. I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review*, 94(4):813–835.

Seely, J. C. (2005). The legacies of transition governments: post-transition dynamics in Benin and Togo. *Democratization*, 12(3):357–377.

Shenton, L. and Bowman, K. (1977). A bivariate model for the distribution of $\sqrt{b1}$ and b2. *Journal of the American Statistical Association*, 72(357):206–211.

The World Bank (2015). World Development Indicators, retrieved 12 March, 2015.

Van de Sijpe, N. (2013). Is foreign aid fungible? Evidence from the education and health sectors. *The World Bank Economic Review*, 27(2):320–356.

# APPENDIX

# A The relative importance of prices and NA data

## A.1 Construction of alternative GDP series

Using the new price data, I compute the GDP at constant 2005 international (PPP) dollars using:

$$y_t = \frac{c_t^{NA}}{PPP_{c,05}} + \frac{g_t^{NA}}{PPP_{g,05}} + \frac{inv_t^{NA}}{PPP_{i,05}} + \frac{x_t^{NA}}{PPP_{gdp,05}} - \frac{im_t^{NA}}{PPP_{gdp,05}}$$

where $y_t$ is GDP, $c_t$ is consumption, $g_t$ is government expenditure, $inv_t$ is investment, $x_t$ are the exports and $im_t$ are the imports at time $t$. The superscript $NA$ means that the data are taken from the national accounts tables underlying PWT6.[16] All of these are absolute values in national currencies, adjusted for inflation.[17] Note that the methodology for inflation adjustments can differ between countries, as these are normally carried out by the national statistical offices. The expenditure category specific PPPs are not explicitly reported in PWT7, so I compute them as

$$PPP_{i,05} = (p_{i,05}^{PWT7}/100) * XRAT_{05}^{PWT6} \tag{3}$$

where the $i = c, g, inv, x, im$, with meanings identical to above. PWT6 uses a country's overall price level of domestic absorption (a weighted average of $p_c, p_i$ and $p_g$) for exports and imports, meaning here that $p_x = p_{im} = p$. $XRAT_{05}^{PWT6}$ is the 2005 exchange rate as reported in PWT6; I do not use the exchange rates underlying PWT7 in order to isolate the effect of the prices. The method used to construct the GDP series based on PWT7 NA data and PWT6 prices is identical, with the only exception that equation 3 becomes obsolete as PPPs are provided in PWT6.[18]

## A.2 Visual inspection for complete sample

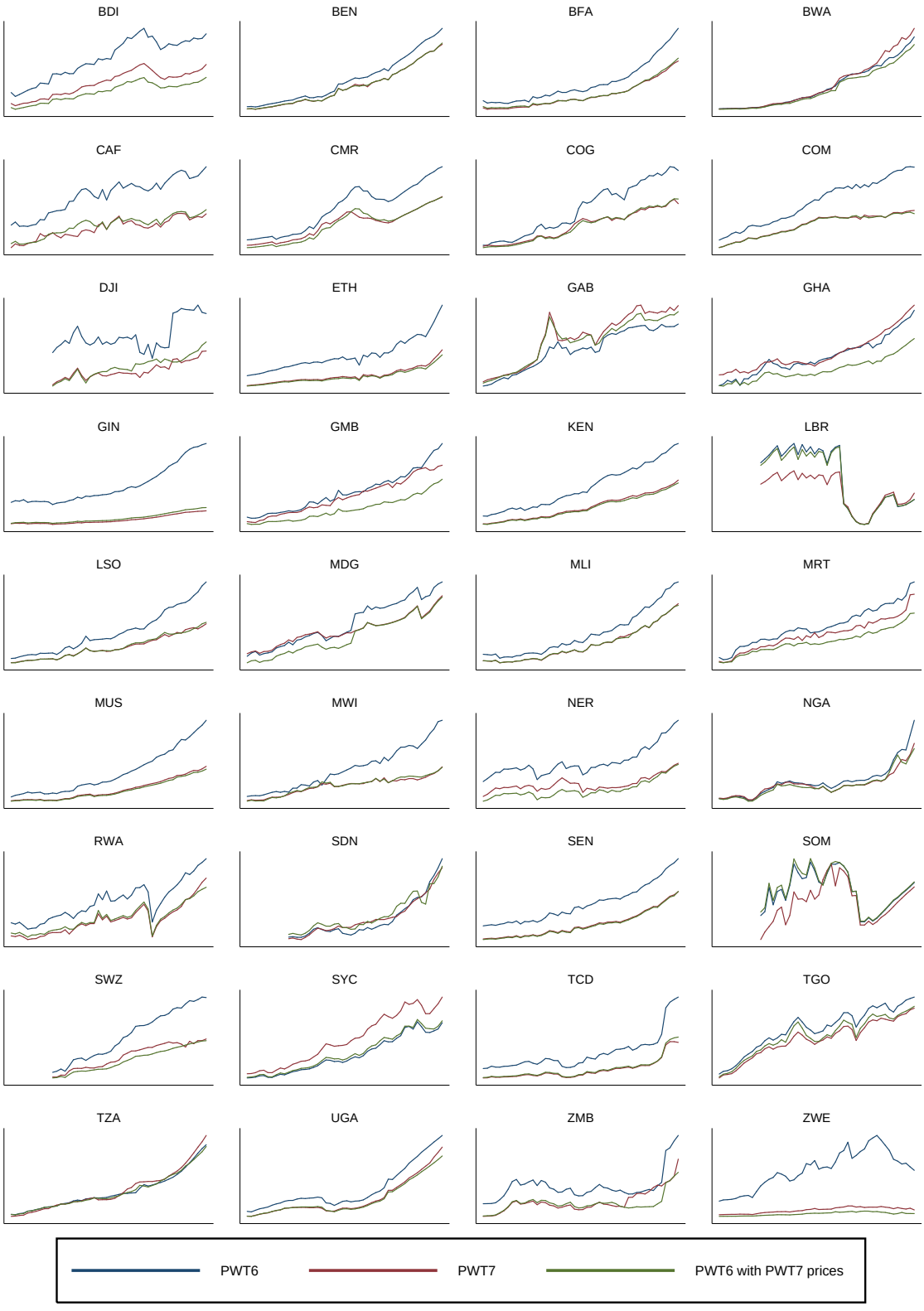
Figure 5 reproduces figure 3 for the entire sample of 36 Sub-Saharan African countries. In almost all countries, the PWT6 series compiled using PWT7 prices matches PWT7 significantly better than the original PWT6 series. Besides Gabon, which is discussed in the main body of the paper, this is particularly pronounced in the cases of Djibouti and Zimbabwe. In these countries, the two versions of the Penn World Table report strikingly different dynamics, judging from visual inspection of the graphs. After adjusting the prices, the growth patterns are largely reconciled, again highlighting the importance of the relative prices attributed to the expenditure shares. Indeed, in Djibouti, the estimated price of consumption has almost doubled between the versions, while that of investment has been more than halved. For Zimbabwe, 2005 price figures were revised upwards by a factor of up to 10 or 20, while the exchange rate has been lowered by a factor of 22 between the two vintages; in fact, the PWT7 documentation mentions the country as a problematic case in terms of price collection because of its high inflation rates (Heston et al., 2012).

[16] Provided by the authors at `http://www.rug.nl/research/ggdc/data/pwt`.

[17] That is, they are in real terms; I avoid using the term in order to avoid confusion, as PWT typically use it to describe PPP adjusted values (made comparable across countries rather than time).

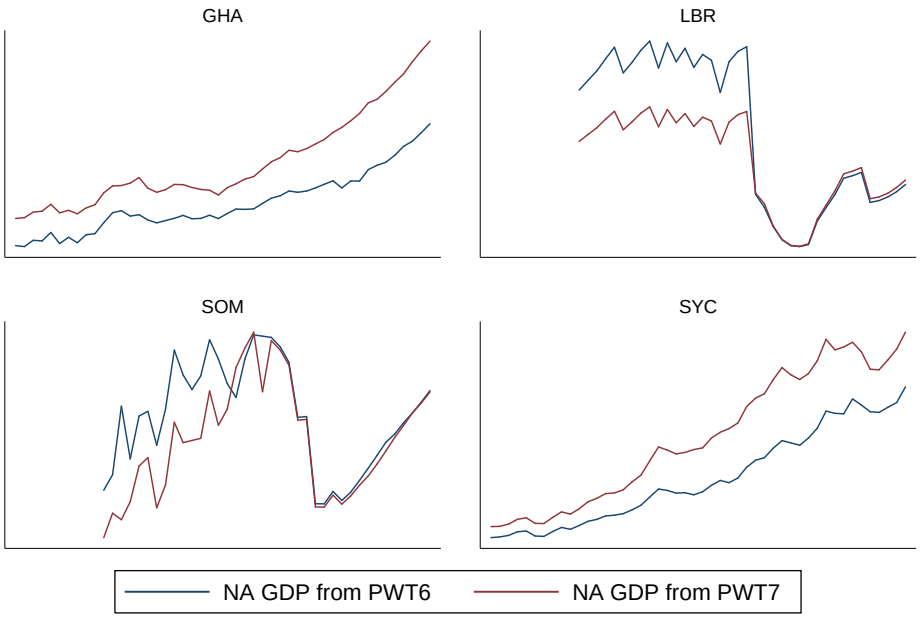
[18] Stata do-files are available from the author upon request

Figure 5: GDP for 36 countries, 1960-2007




In a few countries however, namely Ghana, Liberia, Somalia, and Seychelles, the new prices have little discernible impact on the GDP series. While in all of them there clearly is substantial divergence between the datasets, the relative prices of their GDP components

Figure 6: GDP from underlying NA data, 1960-2007




have remained relatively stable. Instead, their underlying National Accounts data has been strongly revised: Figure 6 juxtaposes the constant price GDP series as reported in the NA datasets underlying PWT6 and PWT7 respectively. These are the figures as reported by the statistical offices, prior to any PPP adjustments or other alterations stemming from the PWT methodology. Comparing these to the corresponding plots on figure 5, it is apparent that the patterns of divergence are very similar.

# B Consistency by country

Table 10: Consistent coefficients by country

| *Country* | *PWT6* | *PWT7* | *PWT8* | *WDI* | *Consistent* |
|---|---|---|---|---|---|
| BDI | 4 | 4 | 2 | - | 10/12 |
| **BEN** | **4** | **2** | **2** | **2** | **10/16** |
| **BFA** | **4** | **4** | **4** | **4** | **16/16** |
| BWA | 1 | 0 | 0 | - | 1/12 |
| CAF | 4 | 4 | 0 | - | 8/12 |
| **CMR** | **4** | **4** | **4** | **4** | **16/16** |
| **COG** | **4** | **4** | **3** | **3** | **14/16** |
| COM | 4 | 0 | 3 | - | 7/12 |
| DJI | 2 | 2 | 0 | - | 4/12 |
| ETH | 4 | 1 | 0 | - | 5/12 |
| **GAB** | **4** | **4** | **4** | **4** | **16/16** |
| GHA | 4 | 2 | 1 | - | 7/12 |
| GIN | 4 | 3 | 4 | - | 11/12 |
| GMB | 4 | 1 | 3 | - | 8/12 |
| **KEN** | **4** | **4** | **4** | **4** | **16/16** |
| LBR | 4 | 4 | 0 | - | 8/12 |
| **LSO** | **4** | **0** | **0** | **0** | **4/16** |
| **MDG** | **4** | **4** | **3** | **3** | **14/16** |
| MLI | 3 | 3 | 3 | - | 9/12 |
| **MRT** | **1** | **0** | **1** | **2** | **4/16** |
| MUS | 4 | 3 | 3 | - | 10/12 |
| MWI | 4 | 3 | 3 | - | 10/12 |
| NER | 4 | 3 | 4 | - | 11/12 |
| NGA | 4 | 2 | 2 | - | 8/12 |
| **RWA** | **3** | **2** | **3** | **2** | **10/16** |
| SDN | -* | 2 | 2 | 3 | 10/16 |
| **SEN** | **3** | **4** | **4** | **1** | **12/16** |
| SOM | 4 | 3 | 0 | - | 7/12 |
| SWZ | 4 | 2 | 2 | - | 8/12 |
| SYC | 3 | 4 | 0 | - | 7/12 |
| TCD | 4 | 4 | 4 | - | 12/12 |
| **TGO** | **2** | **2** | **2** | **1** | **7/16** |
| TZA | 3 | 1 | 1 | - | 5/12 |
| UGA | 4 | 1 | 2 | - | 7/12 |
| ZMB | 3 | 2 | 2 | - | 7/12 |
| ZWE | 4 | 1 | 1 | - | 6/12 |

The table reports the number of coefficients that are consistent with those obtained by JMT under the prior of aid effectiveness, taking into consideration the first and second best choice of rank, using their exact models. The last column reports the total sum of consistent coefficients within each country across all datasets, followed by the number of estimated coefficients. Countries included in the re-specification exercise in section 5.3 are in bold characters.

* JMT use WDI data only in the case of Sudan for reasons of data availability, which is why the corresponding PWT6 figure is left out.

# C Determinants of consistency

In order to make an assessment of the stability of one's results with respect to different datasets without having to run an entire analysis every time, it would be beneficial to have a metric or a set of metrics that have a reasonable predictive power concerning the stability of the results. It is however difficult to discern an obvious relationship between the consistency of the results and the properties of the underlying data. Figures 1 and 2 in section 3.2 can give a first sense of this. Both the countries with the most stable results in the subsequent analysis (top panels) and those with the least stable results (bottom panels) seem to diverge to comparable extents. This is true both for GDP levels and, as illustrated with shares of investment in figure 2, shares of GDP.

Figure 7 abstracts from differences in levels and plots the growth rates of the same countries over the sample period. Again, it is hard to discern any clear patterns, as both sets of countries (the four with the most stable result at the top, the least stable ones at the bottom) exhibit quite volatile and divergent growth rates. Besides the visual inspection of the graphs, I tried to relate a number of measures to the stability of the results. None of them turned out to be a good predictor. These included:

- Variance (alternatively standard deviation) around the average between the series:

$$\sigma^2_{Y,t} = \frac{1}{4} \sum_{i \in PWT6,PWT7,PWT8,WDI} (Y_{i,t} - \frac{\sum_{i \in PWT6,PWT7,PWT8,WDI} Y_{i,t}}{4})^2$$

- As this measure mechanically increases as GDP increases, the same with GDPs normalised around their mean each period (deviation from mean in percent),

$$Y^{norm}_{i,t} = \frac{Y_{i,t}}{\frac{1}{4}\sum_{i \in PWT6,PWT7,PWT8,WDI} Y_{i,t}} * 100$$

  and

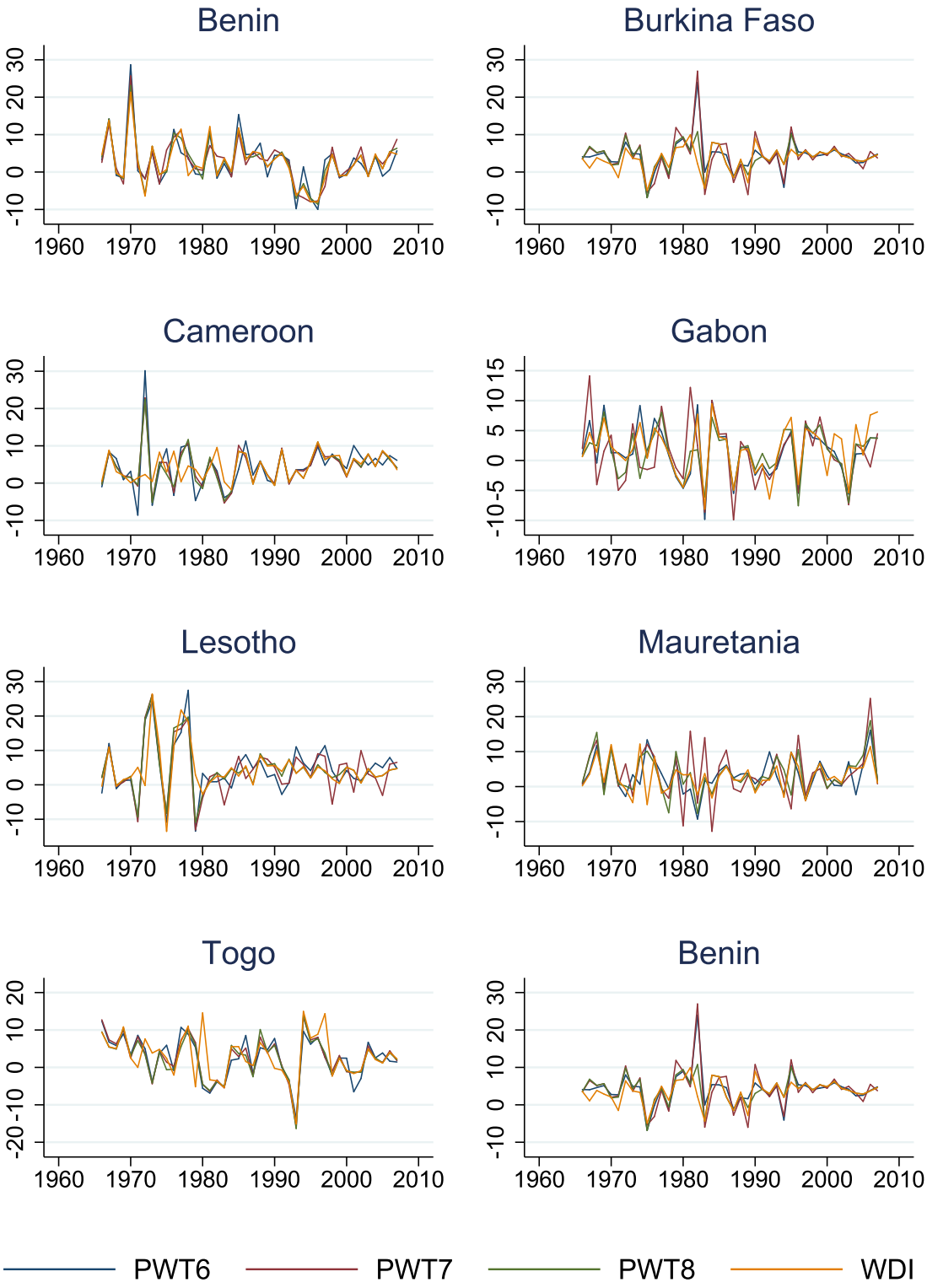
$$\sigma^2_{Y,i,t} = \frac{1}{4} \sum_{i \in PWT6,PWT7,PWT8,WDI} (Y^{norm}_{i,t} - 100)^2$$

  .

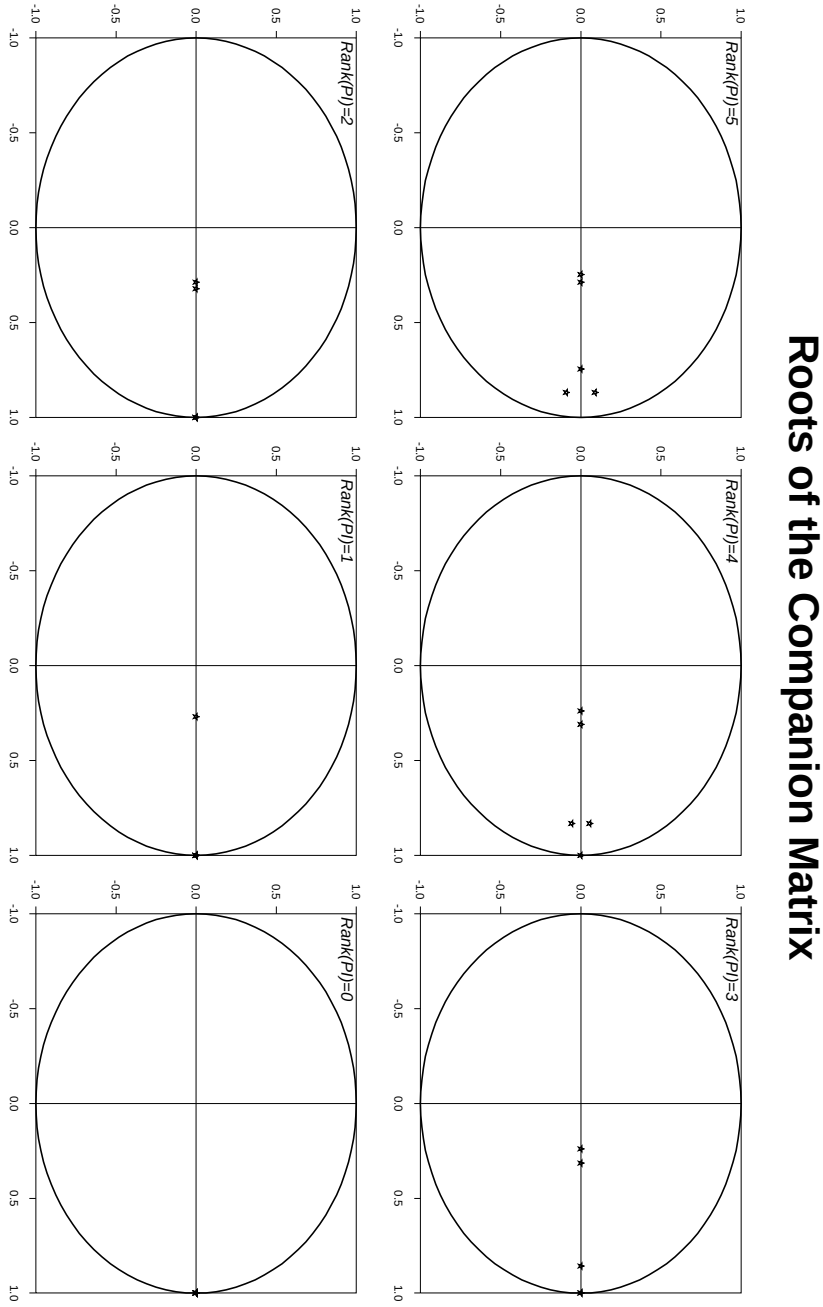
- Variance and standard deviation across series (as above) in terms of growth rates
- Variance and standard deviation as described above for individual expenditure shares.
- Statistical capacity as reported by the World Bank. While this has some power in predicting the consistency of individual series over time, it does not predict the consistency of the results. This is in line with the findings from the previously mentioned measures.

The strongest contributing factor I could identify was not in the data, but in the results. In the present sample, insignificant results were much more likely to persist throughout the datasets: Within the 8 countries that had no single significant coefficient in the first place (PWT6), 80% of the replicated coefficients were also insignificant in the replications. Within all other countries, 50% kept their sign and significance.

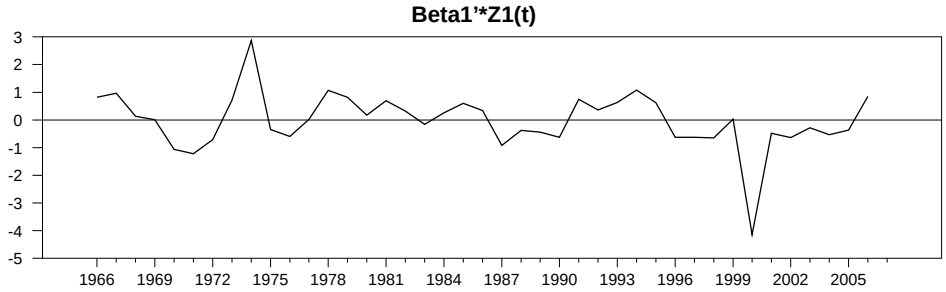
Figure 7: GDP growth rates from four sources

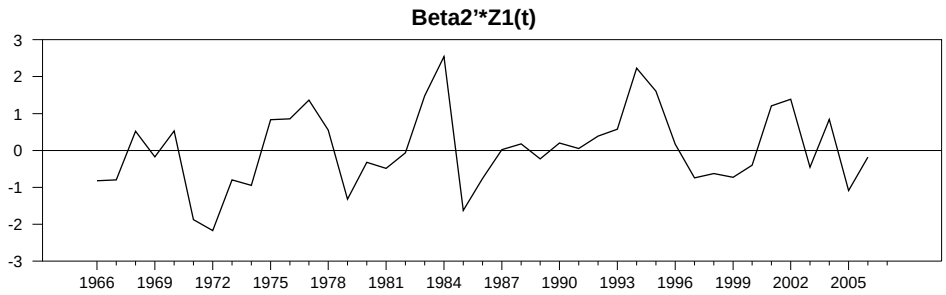


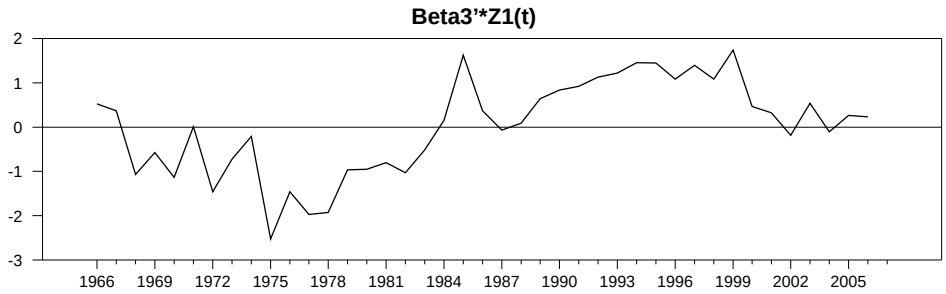
# D Roots of the companion matrix (Burkina Faso, PWT6)

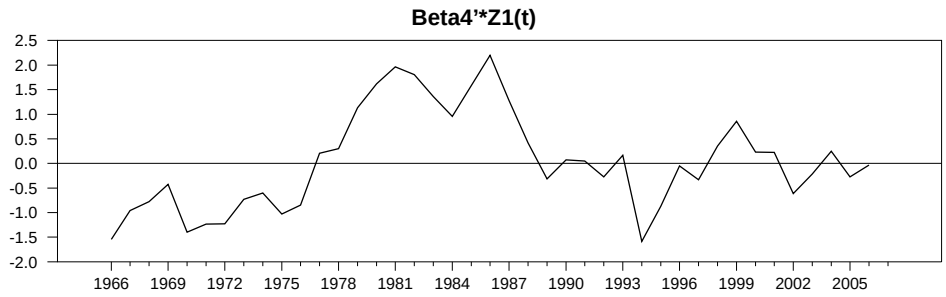


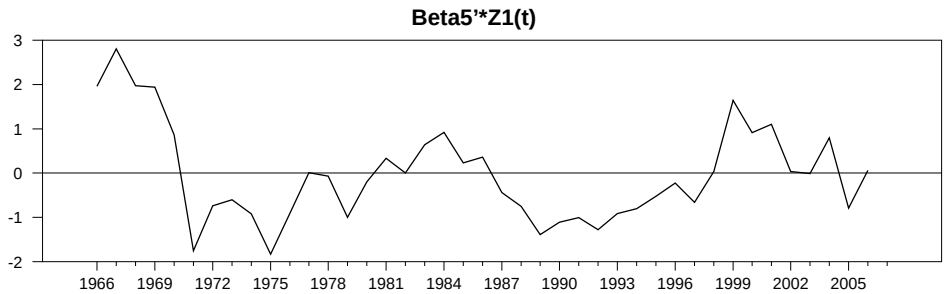
# E Graphs of the cointegrating relations (Burkina Faso, PWT6)



# F T-ratios of the $C_{21}$-Matrix

Table 11: $t$-ratios of the best and second best choice of rank

| | **PWT6** | | **PWT7** | | **PWT8** | | **WDI** | |
|---|---|---|---|---|---|---|---|---|
| | $r^*$ | $r'$ | $r^*$ | $r'$ | $r^*$ | $r'$ | $r^*$ | $r'$ |
| Burkina Faso | -1.03 | -1.08 | -1.08 | -0.35 | -1.03 | -1.08 | 0.65 | -0.24 |
| | 1.15 | -0.13 | 0.68 | -4.00 | 1.15 | -0.13 | -3.81 | -1.52 |
| | -0.26 | -0.43 | -0.29 | 1.78 | -0.26 | -0.43 | 2.36 | 0.72 |
| | 0.16 | 0.89 | -0.41 | 1.70 | 0.16 | 0.89 | -2.37 | 1.68 |
| Cameroon | -0.55 | -1.19 | -0.91 | -0.93 | -0.89 | -0.68 | -1.24 | -3.01 |
| | 0.62 | -0.30 | -0.22 | -1.12 | -0.21 | -1.20 | 0.09 | 0.12 |
| | -0.15 | -0.59 | -0.82 | -0.26 | -0.80 | -0.03 | -0.28 | -2.07 |
| | -0.45 | -3.85 | -0.61 | -0.54 | -0.27 | 0.73 | -1.48 | -0.78 |
| Gabon | 0.23 | 1.02 | -0.41 | 0.14 | 1.89 | 1.81 | -0.50 | 0.65 |
| | -0.01 | 1.02 | -0.41 | 0.19 | 1.88 | 1.81 | -0.59 | 0.57 |
| | -2.80 | 1.02 | -0.42 | -0.02 | 0.68 | 1.22 | -0.88 | -0.69 |
| | 1.18 | 1.02 | -0.42 | 0.12 | 1.32 | 1.55 | -0.86 | 0.77 |
| Kenya | 4.03 | 1.44 | 2.54 | 2.84 | 2.69 | 3.02 | 4.75 | 2.23 |
| | 3.38 | 3.09 | 2.01 | 1.88 | 2.46 | 1.82 | 4.53 | 2.23 |
| | 3.83 | 1.80 | 2.59 | 2.96 | 2.72 | 3.04 | 4.41 | 2.23 |
| | 0.67 | 1.05 | -1.13 | -0.54 | -0.18 | -0.15 | 1.69 | -2.23 |
| Congo, Rep. | -0.09 | 1.36 | 0.36 | -0.55 | -0.36 | -0.32 | -0.13 | -0.10 |
| | 1.16 | 1.19 | 0.84 | 1.40 | 0.31 | 1.09 | -0.13 | -0.12 |
| | -0.48 | 0.17 | 0.84 | -0.94 | -0.24 | -0.91 | -0.13 | -0.06 |
| | 0.93 | 1.21 | 0.68 | 0.52 | -0.14 | 0.34 | -0.13 | -0.11 |
| Madagascar | 2.47 | 0.66 | -1.37 | 0.00 | 0.20 | 1.97 | -0.75 | -0.80 |
| | 2.47 | -0.63 | -1.37 | -2.16 | -0.71 | 16.62 | 0.70 | 0.42 |
| | 2.47 | 4.90 | 1.37 | 3.05 | 2.18 | 2.04 | -1.19 | -0.46 |
| | -2.47 | -6.48 | -1.37 | -0.98 | -0.54 | -1.07 | 0.90 | -1.06 |
| Rwanda | 1.10 | -5.09 | 1.65 | 3.23 | 2.97 | 3.04 | 4.33 | 4.27 |
| | 2.10 | -3.58 | 1.84 | 2.29 | 1.84 | 2.17 | 4.07 | 1.64 |
| | 3.42 | 1.08 | 2.40 | 1.59 | 2.59 | 1.46 | 3.14 | 3.57 |
| | -1.84 | -0.95 | -0.23 | 0.21 | -0.07 | 0.27 | 1.10 | 1.59 |
| Senegal | 2.01 | 4.16 | 1.54 | 3.05 | 1.30 | 4.01 | 0.47 | -0.16 |
| | 0.44 | 2.74 | -0.93 | 0.97 | 0.58 | 3.09 | 1.30 | 1.17 |
| | 2.06 | 2.25 | 2.65 | 3.42 | 1.98 | 2.87 | 3.39 | 3.39 |
| | 2.19 | 3.99 | 2.86 | 3.56 | 1.93 | 3.58 | 1.10 | 1.74 |
| Benin | 1.15 | -2.37 | -0.63 | 6.70 | -0.89 | 2.31 | -0.75 | -1.81 |
| | 3.05 | -2.37 | -4.26 | 12.78 | 0.15 | 9.97 | 0.70 | -4.15 |
| | 3.74 | 2.37 | 4.19 | 4.58 | 2.48 | 2.43 | -1.19 | 0.76 |
| | -2.84 | -2.37 | -1.03 | 6.43 | -2.10 | -4.38 | 0.90 | 2.54 |

Table 11: Table 11 (continued)

| | **PWT6** | | **PWT7** | | **PWT8** | | **WDI** | |
|---|---|---|---|---|---|---|---|---|
| Lesotho | 2.34 | 5.11 | -5.94 | -11.51 | -0.12 | 0.83 | 1.09 | -7.04 |
| | 2.22 | 2.21 | 1.56 | -1.57 | 0.93 | 3.72 | 1.10 | -2.94 |
| | 3.28 | 3.46 | -6.90 | -6.15 | -2.55 | -7.35 | 1.47 | -2.50 |
| | 2.27 | 0.02 | -1.08 | -2.43 | -7.02 | 0.49 | -1.95 | -1.00 |
| Mauritania | -1.30 | 0.00 | 1.85 | -3.64 | -0.93 | -2.97 | -5.20 | -0.31 |
| | 2.17 | -0.26 | 0.18 | -2.91 | 0.75 | -0.02 | -0.75 | -26.12 |
| | 0.58 | 0.06 | 4.51 | -0.51 | 2.93 | 4.80 | 0.27 | 0.91 |
| | 2.84 | 0.08 | 0.97 | 3.06 | 0.90 | 0.41 | 2.38 | -1.78 |
| Togo | 3.64 | 1.99 | 11.16 | 2.47 | 11.63 | 1.49 | 4.22 | 3.66 |
| | -0.94 | 1.73 | -3.62 | -4.98 | -3.55 | -4.81 | 2.29 | 0.81 |
| | 4.88 | -1.41 | 5.83 | 1.63 | 4.26 | 1.04 | 6.45 | 6.61 |
| | 2.22 | 3.73 | 1.11 | 1.13 | 0.37 | 0.58 | 0.88 | 0.68 |

Notes: $r^*$ reports the $t$-ratios for the variables in the second column under the preferred rank specification, $r'$ those obtained under the second best choice of rank.

Source: Author's calculations.