Discussion Papers
Department of Economics
University of Copenhagen

No. 14-28

Structural transformation in the 20th century:
A new database on agricultural employment around the world

Asger Moll Wingender

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

Structural transformation in the 20th century: A new database on agricultural employment around the world

Asger Moll Wingender*

University of Copenhagen

December 2, 2014

Abstract

Many empirical questions about economic growth and development are left open due to the lack of long time series of reliable GDP estimates. The share of the labor force employed in agriculture can fill this gap. Agricultural employment shares are highly correlated with GDP per capita, less prone to measurement errors, and data are available for longer periods than existing GDP estimates.

This paper describes a new database on agricultural employment covering 169 countries for the period 1900-2010. Some of the many potential uses of the data are discussed.

KEYWORDS: Economic growth, structural transformation, agricultural employment.
JEL: O1, O4.

*I am grateful to Carl-Johan Dalgaard for comments and suggestions, and to David Good for sharing his census data from the Habsburg Empire.

1 Introduction

Why are some countries rich and some countries poor? It is arguably one of the most important question in macroeconomics, and it often shows up in introductions to papers on economic growth. No answer to the question is forthcoming without a reliable yard stick for measuring income differences between countries. Gross Domestic Product (GDP) per capita is an obvious choice of yard stick, as GDP is designed to be an empirical counterpart to aggregate output in a macroeconomic production function. Indeed, the aim of an ever growing empirical literature is to explain differences in growth rates or levels of GDP using regression analysis or accounting techniques.[1]

GDP has its shortcomings, however. It is a statistical concept rather than an observable quantity, and extensive information on production and prices, as well as complicated statistical methods, are required to estimate GDP. Many developing countries do not have the necessary statistical capacity to do so, and their GDP estimates are consequently unreliable.[2] The unreliability is illustrated by the recent upward revisions to GDP in Ghana and Nigeria of 62 and 89 percent respectively. Both revisions followed a change of base year for the price indecies used to calculate real GDP. As discussed in Section 6.5, equally big revisions in many poor and middle income countries have been caused by base year changes for the purchasing power parities (PPPs) used in international comparisons.

Another issue is that the modern concept of GDP was not formalized until the first System of National Accounts was published by the United Nations in 1953, and many countries did not publish official GDP estimates until decades later. Historical GDP estimates do exist for some countries thanks to the valuable work of economic historians, notably Angus Maddison and other researchers affiliated with University of Groningen Growth and Development Centre (GGDC). But the lack of raw historical data on prices and production means that such GDP estimates often suffer from high margins of error, not unlike data from present day Africa.

[1] Surveys of the two literatures can be found in Barro (1996) and Caselli (2005) respectively.

[2] See Jerven (2013) for a book-length survey of the quality of national accounts in sub-Saharan Africa.

While much has been learned from the empirical growth literature, no clear prescription for spurring growth in developing countries has emerged. It is certainly plausible that no such prescription exists. But missing and unreliable GDP data may also make it hard to distinguish useful policies from useless ones.[3]

That is the motivation behind the data collecting project underlying this paper. As an alternative to the GDP data, I have compiled a comprehensive database of agricultural employment shares for the period 1900-2010. Agricultural employment is closely related to national income. Poor countries tend to have almost the entire labor force employed in the fields, whereas agriculture is a negligible source of employment in rich countries. The database, available online, covers 169 of the 177 independent countries that had more than 250,000 inhabitants in 2010.[4]

I use urbanization rates to extend the database to periods where no employment data are available. Agriculture is, almost by definition, a rural activity, whereas cities are more favorable for most other economic activities. There is a close and stable empirical relationship between agricultural employment shares and urbanization rates, and I show that this relationship can be used to accurately estimate agricultural employment shares.

Agricultural employment shares are useful as an alternative to GDP data for several reasons. First and foremost, the share of the labor force engaged in agriculture is closely related to productivity and national income through Engel's Law. A subsistence food requirement increases the consumption share of agricultural goods in countries with low productivty levels, and hence low national income. To satisfy the high relative demand for food, more workers are needed in agriculture in poor countries than in countries with high productivity levels. An alternative would be to import the required food, but poor countries rarely do so on a sufficient scale. Low income levels are therefore reflected in sectoral employment rather than international trade. Empirically, the correlation between GDP and agricultural employment is 0.9 in a cross section of 158 countries in 2000, and, as shown in Section 6, the relationship has

[3]This point is forecefully made by Ciccone and Jarociński (2010).

[4]The database is available at https://sites.google.com/site/asgerwingender/.

Table 1: Earliest census with employment data

Europe and North America		Other regions	
Country	**Year**	**Country**	**Year**
Finland	1774	Brazil	1872
Iceland	1801	Japan	1872
Norway	1801	Argentina	1895
United States	1820	Mexico	1900
United Kingdom	1841	India	1901
Belgium	1846	Taiwan	1905
Netherlands	1849	Indonesia	1905
Denmark	1850	Egypt	1907
France	1856	South Africa	1911

Sources: Mitchell (1993, 1998a,b), Minnesota Population Center (2008).

been stable over time despite rapid globalization.[5] The link between agricultural employment and income is a well-known stylized fact of development economics, and have been analyzed extensively. Useful overviews of the literature cen be found in Gollin (2010) and Herrendorf *et al.* (2014).

Agricultural employment shares have additional advantages over GDP data. They are much simpler to measure, and consequently less prone to measurement errors. Moreover, governments have usually carried out censuses or labor force surveys before they were able to accurately calculate GDP, or before GDP was even invented. Agricultural employment shares are, for instance, readily available for a number of countries from the 19th century and onwards. Examples are shown in Table 1.

The present study is not the first to be motivated by lacking and uncertain GDP estimates. Other researchers have looked for alternative income measures for the same reasons. Chen and Nordhaus (2011, 2014) and Henderson *et al.* (2012) use the intensity of night lights measured from space by satellites as a proxy for GDP. IMF (2006) uses growth in electricity consumption to show that GDP growth in Jamaica was probably 3.1 percent per annum from 1991-2000 rather than the official estimate of 0.3 percent. Young (2012) shows that the living standards in

[5]Sources: Penn World Table 8.0 compiled by Feenstra *et al.* (2013), and the agricultural employment share data documented in this paper.

sub-Saharan Africa, according to consumption data from the Demographic and Health surveys, have grown by three-and-a-half to four times faster than GDP per capita, indicating that GDP may be underestimated. Consistent with this result, I show in Section 6 that GDP in sub-Saharan Africa also appears to be underestimated when it is compared to the region's falling agricultural employment shares. To demonstrate that it is the African employment data that give the more accurate picture of income levels, I show that income levels predicted by lights from space correspond to the ones implied by agricultural employment shares rather than the official GDP estimates. Lights data are not available before 1992, but it seems reasonable to assume that result would be similar in earlier periods if data had existed.

Other cases of measurement errors are also visible when comparing the GDP data to agricultural employment shares. An example is the well-know overestimation of GDP in the USSR and Eastern Europe during communism. To go beyond case studies, I show in Section 6.5 that agricultural employment data can predict revisions to GDP following changes of base year for the PPP calculations. By implication, agricultural employment shares contain information on true income levels not fully reflected in the PPP adjusted GDP data.

This paper is mostly concerned with the relationship between agricultural employment and national income. But beyond being an alternative to GDP in, *e.g.*, cross country growth regressions, agricultural employment shares are useful for studying many issues not necessarily related to national income. Changed sectoral employment patterns may, for instance, affect fertility, mortality, institutions, and cultural and social norms. The database can also be used to test theoretical models of structural change, investigate the spread of industrialization, or to study the mechanics of dual economies (*i.e.*, economies where large unproductive agricultural sectors coexists alongside small and productive modern sectors). I leave these possibilities as topics for future research.

The paper is structured as follows. The theoretical and empirical links between agricultural employment and GDP are reviewed in Section 2, and measurement errors in the two variables are discussed. The sources of employment data are described in Section 3. There are a number of existing databases, notably the ones maintained by The International Labor Organization

(ILO), GGDC, Oxford Latin American Economic History Database, OECD, and the International Historical Statistics by Mitchell (1993, 1998a,b). I merge these databases, and extend the resulting data set with information from numerous other sources, including data collected from national statistical offices, various issues of the Yearbook of the League of Nations, and research by economic historians.

Urbanization data are for most countries available in earlier periods than employment data. In Section 4, I describe how I use urbanization rates to estimate agricultural employment in periods when employment is unobserved. The coverage of the resulting data set is described in Section 5, and, in section 6, I compare the evolution of agricultural employment in the 20th century to the evolution of GDP per capita. Section 7 concludes by discussing potential uses of the database, and avenues for further research.

2 Agricultural employment and national income

The rationale for using agricultural employment shares to study economic development, made in the introduction, is spelled out in further details in this section. Agriculture (including fishing) was from the onset of the Neolithic Revolution to the eve of the Industrial Revolution the dominant sector around the globe. By the early 19th century, manufacturing had overtaken agriculture in the United Kingdom, from where the Industrial Revolution gradually spread to other parts of Europe and the English speaking world, albeit delayed by almost a century. The transition out of agriculture was accompanied by economic growth unprecedented in human history. Productivity increases outpaced fertility to an extent that fewer workers in the fields were needed to feed the population.

This development can be formalized in a very simple model. Let preferences be of the Stone-Geary variety, such that the utility function of the representative individual takes the form:

$$u\left(c_a, c_n\right) = \left\{ \begin{array}{cc} c_a & if\ c_a \leq \bar{c}_a \\ \ln\left(c_n\right) + \bar{c}_a & if\ c_a > \bar{c}_a \end{array} \right\},$$

where c_a is consumption of agricultural goods (food), and c_n is consumption of nonagricultural goods.[6] An extreme version of Engel's Law holds in this formulation of preferences. Consumers only care about their calorie intake when food consumption is below the satiation point $\bar{c}_a$. Above the satiation point, only nonagricultural goods increase utility.

Labor is the only input in production in the two sectors, and output is proportional to a common productivity level Z, which include technology, physical capital, human capital etc. Agricultural production per capita is consequently given by $y_a = Z \cdot AES$, where AES is the agricultural employment share. Nonagricultural production per capita is similarly given by $y_n = Z \cdot (1 - AES)$.

Let the economy be closed such that $c_a = y_a$. It follows that $AES = \min\left\{\frac{\bar{c}_a}{Z}, 1\right\}$, so countries with low productivity levels have large fractions of their workforce employed in agriculture. The productivity level, Z, is the only source of possible variation in aggregate income across countries, and agricultural employment shares are therefore proportional to GDP per capita.

The reality is, of course, infinitely more complicated than the model above. But the relationship between agricultural employment and income is also a feature of more realistic models. Recent examples include Lucas (2009), Lagakos and Waugh (2013), Gollin and Rogerson (2014) and Wingender (forthcoming).

It is no coincidence that theoretical two-sector models predict a high correlation between agricultural employment shares and income. They are build to match that relationship, as it is one of the most robust stylized empirical facts of economic development. It is illustrated in Figure 1, which shows log GDP per capita as a function of the agricultural employment share for a cross section of countries.

The regression line in Figure 1 has an R^2 of 0.8, and as shown in Section 6, the goodness of

[6] This formulation of preferences is used in Laitner et al (2000) and Gollin et al (2002).

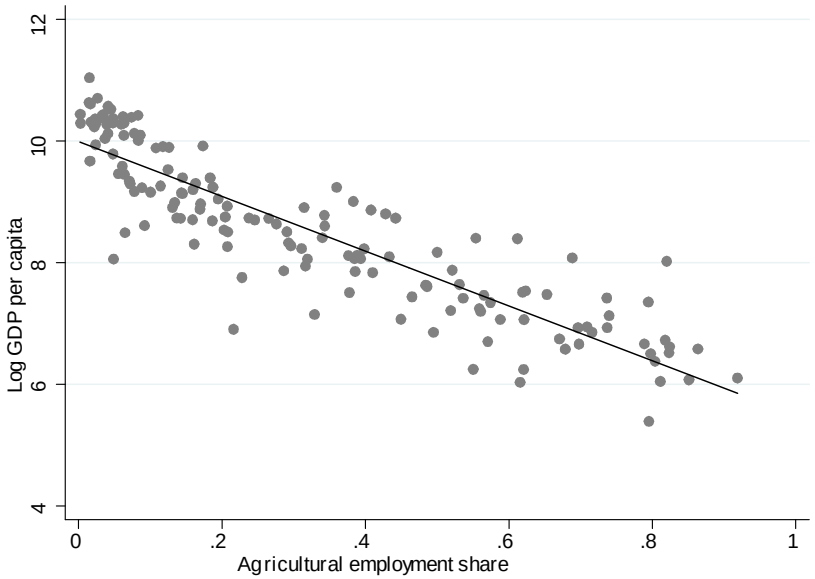

Figure 1: Agricultural employment shares and log GDP per capita 2000

fit is similar in earlier periods. While the observed relationship between GDP and agricultural employment is close, it is not perfect due to measurement errors, and errors introduced by the simplicity of the model. I discuss errors in details in the next two subsections. One conclusion that emerges is that measurement errors in GDP, while hard to quantify, are likely to be substantial. The implication is that agricultural employment shares are even better predictors of unobserved true income than of observed GDP per capita.

2.1 Measurement errors

GDP is a complex statistical concept. To compute GDP, detailed statistics on production and prices are required, as well as a sizeable number of statisticians and computers to process the raw data. Neither are available in developing countries. Raw data are often non-existing, as data collection is costly, and little is known about economic activity in the large informal sector. Furthermore, limited resources are allocated to process the raw data that do exist. Jerven (2013) describes how one single person in the Central Statistical Office in Zambia was responsible for calculating all the income and growth statistics when he visited in the end of

the 2000s. Visits by Jerven to other statistical offices in Sub-Saharan countries showed that Zambia is by no means unique in this respect.

Lack of data on prices is especially a worry. The recent large revisions to GDP in Ghana and Nigeria, mentioned in the Introduction, were caused by changes in the national base year used to calculate real GDP in constant prices. Both countries changed their base year to 2008, from 1993 in Ghana and 1990 in Nigeria. GDP calculated using constant prices is biased if relative prices move.[7] Price movements are often dramatic in economies where structural change happens quickly, as in Ghana and Nigeria, and the measurement error accumulates the further in time from the base year GDP is estimated.

Major GDP revisions following base year changes are no longer common in the developed world where chain-weighting is used to calculate real GDP. Chain-weighting requires detailed annual data on the price structure of the economy, and the method is not feasible in developing countries where price surveys are not carried out on a regular basis. The constant price approach is used instead. But unlike Ghana and Nigeria, many developing countries still use base years from the 1990s or 1980s, with huge implications for the accuracy of their official estimates of real GDP.

GDP levels need to be adjusted for differences in national prices to be comparable across countries. To this end, most researchers use PPP adjustment factors published by the International Comparison Program (ICP). The PPPs are calculated in a base year, and extrapolated using national price indecies. Changes in base year for the PPPs often cause significant revisions to relative GDP levels for the same reason as changes of base year in the national price data. Moreover, 22 countries did not submit price data for one or more component of the national accounts in the 2011 round of the ICP program, and their PPPs therefore contain imputed or estimated values. I discuss PPPs in further details in Section 6.5 where I show that agricultural employment shares can be used to detect measurement errors in the PPPs.

Historical GDP data suffer from many of the same problems as GDP data from developing

[7] This is known as the Gershenkron effect. The sensitivity of GDP to the choice of base year is explained in details in Nuxoll (1994).

countries today. Countries that are currently developed, were relatively poor a century ago, and did not systematically collect data on production and prices. Moreover, the concept of GDP was developed in the middle of the 20th century, and adopted decades later by many countries. Data collection in earlier periods was not aimed at constructing GDP, and historical estimates of GDP therefore contain a mixture of actual observations, estimates, and guesswork.

Knowledge about prices prior to World War II is, for instance, limited in all but a few countries. As in developing countries today, historical GDP numbers are based on constant prices with distant base years rather than chain-weighting. An exception is the United States. When the Bureau of Economic Analysis in the United States recalculated the official GDP estimates using chain weights rather than constant prices, the average annual growth rate between 1929 and 1950 increased from 2.6 percent to 3.5 percent. By implication, the initial level of GDP was lower. A consequence of the revision is, as Maddison (2003) points out, that labor productivity in the United States was lower than in the United Kingdom in 1913. That amounts to a major reinterpretation of economic history. It is, however, not clear how much the GDP level in the United Kingdom would change if it was recalculated using chain weights, so the comparison makes little sense.

Many other criticisms have been raised against GDP. Nordhaus (1996) argues that the gains from technological breakthroughs are not fully captured by GDP, and gives lighting technology as an example. Other authors complain that both levels and growth rates of GDP differ widely across data sources (*e.g.*, Maddison (2010), Penn World Tables, World Development Indicators or national statistics offices), and even from update to update of a given data source.[8] For example, Ciccone and Jarociński (2010) demonstrate that the Penn World Table version 6.2 income estimates leads to substantial changes regarding the role of government, international trade, demography, and geography than the income estimates in version 6.1.

Agricultural employment is an observable quantity that is relatively easy to define and measure. The measurement problem in employment data is consequently smaller than in GDP

[8] *E.g.*, Breton (2012), Johnson *et al.* (2013), Jerven (2013) and Deaton and Aten (2014).

data. Of course, measurement errors are still present in the employment data. Census data may be incomplete, and surveys unrepresentative. But such measurement errors are usually easy to identify by looking at the meta data, which state if some regions or population groups (*e.g.*, self employed or women) are omitted. Such observations can in practice be adjusted or discarded, depending on the nature of the problem.

2.2 Model misspecification errors

Even with accurate measurement, the link between agricultural employment and income is unlikely to be perfect. It breaks down when the national income level is sufficiently high, as rich countries continue to grow after the share of workers engaged in agriculture has fallen to almost zero. Present day differences between agricultural employment shares in Germany, Sweden and the United Kingdom tell us little about relative income in the three countries. But the negligible agricultural employment shares in the three countries tell us that they have fully completed the transition out of agriculture, which an important characteristic when studying comparative development.

International trade may alter the link between agriculture and national income, as trade makes it possible for poor countries to fulfill their subsistence needs by exhanging nonagricultural goods for food on the international markets. They will use this option if they have a comparative advantage in nonagriculture, if transportation costs are sufficiently low, and if the right kind of trade policies are in place. Whether that is the case, and whether international trade have changed the relationship between income and agricultural employment shares over time are consequently empirical questions. There is some evidence that developing countries increasingly tend to be net importers of food, but the traded quantities are not substantial: Sub-Saharan Africa, the least self-sufficient region in the world, produce agricultural goods that covers 85-90 percent of the calorie intake of its population.[9] Moreover, as I show in Section 6, the relationship between measured GDP and agricultural employment shares has actually

[9] Source: FAO (2012).

been stable over the last century.

While international trade do not introduce any systematic errors, it may exacerbate country specific idiosyncrasies. A country with a low population density, and plenty of fertile land is likely to have many people employed in agriculture if it is able to sell crops on the international markets. The United States was an example of this until the second half of the 20th century.

Very unequal societies, where the elite depend on resource rents may also be able to attain relatively high levels of national income even if the majority of the population is employed in subsistence farming. Equatorial Guinea, one of the biggest oil producers in Africa, is an example. It has an agricultural employment share of roughly 60 percent, but, according to some international comparisons, a GDP per capita comparable to Southern Europe.[10] It is arguably the agricultural employment share that gives the most accurate picture of development in Equatorial Guinea.

The sources of error described above should be kept in mind when employment data are used to analyze economic development and structural change, but they do not change the fact that agricultural employment shares are powerful predictors of national income. On that note, the remainder of this paper is devoted to a description of the database on agricultural employment shares I have compiled.

3 Employment data

I outline the available sources of agricultural employment data, and the strategy I use to merge them, in this section. The data appendix provide more details, as does the additional data documentation in Wingender (2014).

I limit the database to countries that were fully independent, and had more than 250,000 inhabitants in 2010. Some of these were parts of larger entities in the earlier periods (*e.g.*,

[10] Source: World Development Indicators. Penn World Table 8.0 puts the GDP per capita level of Equatorial Guinea substantially lower, and roughly equal to that of Columbia. But it is stil substantially higher than what indicated by the agricultural employment share.

the USSR and Yugoslavia), but I report agricultural employment shares based on present day borders. Consistent with many of the data sources I use, individuals employed in fishing, hunting or forestry are categorized as agricultural workers.

The raw data for calculating agricultural employment shares mstly come from national population censuses, household or labor force surveys. Collecting these from national statistics offices is an immense task. Fortunately, much of the information is already provided by the international databases listed in Table 2. I use data from all of them. The exception is the database provided by the Food and Agriculture Organization of the United Nations (FAO) for reasons explained below.

The databases do not contain observations from all countries in the regions they cover, nor do they provide observations for all years. When information is missing (or unreliable), I augment the data with information from other sources. For the recent decades, that is mostly done by tracking down the numbers on the websites of the respective national statistical offices. I have done so for Albania, Angola, Brunei Darussalam, Cabo Verde, Chad, Côte d'Ivoire, Lao, Lebanon, Lesotho, Kosovo, Madagascar, Maldives, Nepal, Nigeria, Solomon Islands, Timor-Leste, Viet Nam and Zimbabwe. Further, I have calculated agricultural employment shares from micro-level census data obtained from the IPUMS-International database for Brazil, Fiji, Guinea, Haiti, Peru, South Sudan, Uganda, Burkina Faso and Uruguay.[11] Data for the USSR republics are obtained from Easterly and Fischer (1995). For years prior to World War II, I rely on various issues of the Statistical Yearbook of the League of Nations, additional historical census data, and research by economic historians. The full list of sources is available in the data appendix.

3.1 Cleaning and merging the databases

Two data source sometimes disagree on the level of agricultural employment for a given country in a given year. The result of a labor force survey may, for instance, differ from the result

[11] The database is compiled by Sobek *et al.* (2013).

Table 2: Employment databases

Database	Coverage	Period	Type
FAO	World	1980-present	Observed and estimates
GGDC EU KLEMS	Europe	1950-present	Observed and estimates
GGDC Africa Sector Database	Africa	1960-present	Observed and estimates
GGDC 10-sector Database	Asia, Latin America and OECD	1990-present	Observed
ILO (KLIM and Laborstat)	World	1969-present	Observed
International Historical Statistics*	World	1800-present	Observed
OECD	OECD	1990-present	Observed
Oxford Latin America Economic History Database	Latin America	1870-present	Observed

*Notes: *Mitchell (1993, 1998a,b)*

of a census. Sometimes the reason for the disagreement can be found in the meta data. A common cause of discrepancies is differences in geographical coverage. Labor force surveys are in some cases only carried out in urban areas, and they consequently understate agricultural employment. The same is true when self-employed are excluded from the survey. Such observations, which the meta data allows me to identify as inconsistent with the remaining data, are removed from the data set. Observations that are clearly outliers are similarly removed if the meta data are missing.

Another cause of discrepancies is that some data sources report estimated rather than observed agricultural employment shares. As a rule of thumb, I discard all observations that are based on extrapolation or model estimates, even if no actual observation is available.[12] I do so, as analyses based on such data risk getting into circular arguments, where assumptions about economic development are tested using data generated from assumptions about economic development.

In practice, I drop a number of observations in the GGDC Africa Sector Database that are estimated based on aggregate productivity growth as implied by the national accounts. A large number of observations in the FAO data set are extrapolated based on a fitted logistic growth

[12]I make one exception by including the EU-KLEMS data set, which are underpinned by sufficiently many actual observations to be considered accurate. Hungary and Austria are exceptions, and EU-KLEMS is not used for these two countries.

path. Moreover, data for a number of countries in the FAO database are pure estimates, as the do not participate in the FAO agricultural census program, and no other sources of employment data exists.[13] It is in the data set not possible to distinguish actual observations from estimates, and I therefore disregard the FAO data set completely.

The remaining observations are broadly in agreement about agricultural employment, and the different data sources can therefore be merged seamlessly. The few exceptions where additional adjustments are needed are described in Wingender (2014).

4 Extending the data set using urbanization rates

Most countries currently collect employment data, but that has not always been the case. The number of countries with employment data is around 90 in the 1950s, down from 169 in the 2000s. Fortunately, urbanization rates can be used to estimate agricultural employment shares in periods when no employment data exist. Agriculture is, almost by definition, a rural activity, whereas cities are more favorable for most other economic activities. There are exceptions to this rule, such as mining, but we should nonetheless expect urbanization to be lower in countries with a high employment share in agriculture.

The expectation is confirmed by the data. Figure 2 shows the close empirical relationship between agricultural employment shares and urbanization in 1970, one of the earliest year with a substantial number of observations of both variables. The estimated parameters for the regression line are shown in Table 3. The table also reports coefficients for regressions made for a pooled sample of all years, as well as for the individual years 1950 and 2000. The estimates seem plausible, as they imply that a country with little urbanization will have all of its population engaged in agriculture ($AES_{i,t} = \frac{-0.81}{0.82} = 0.99$ in the pooled sample). Moreover,

[13] The accuracy of these estimates are questionable. For example, FAO reports an agricultural employment share of around 70 percent in Djibouti. Yet, roughly 75 percent of the population of Djibouti resides in the capital, Djibouti city, and an establishment survey show that there were only 1,690 agricultural holdings in the country in 2006/2007.

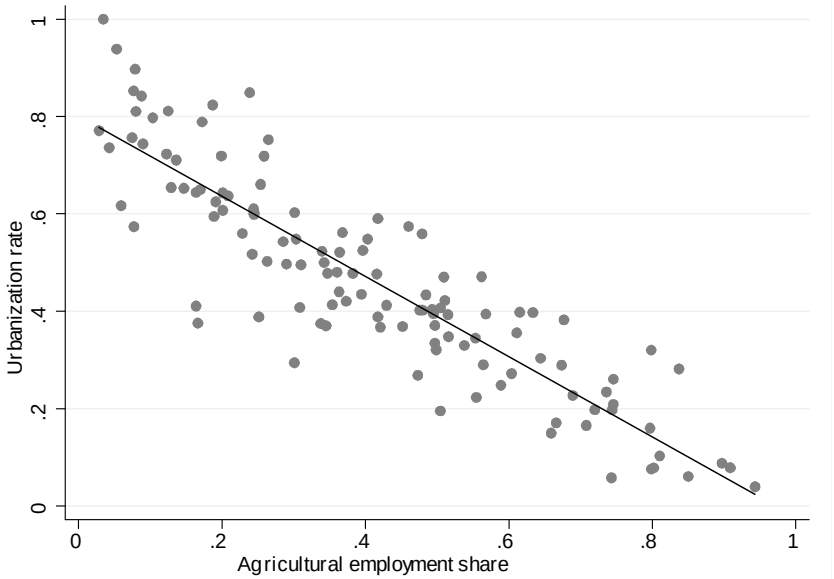

Figure 2: Urbanization and agricultural employment 1970

Table 3: Urbanization and agricultural employment

Sample	**Pooled sample**	**2000**	**1970**	**1950**
Intercept	0.81 (0.00)	0.80 (0.02)	0.80 (0.02)	0.82 (0.03)
Slope (coefficient on AES)	−0.82 (0.00)	−0.80 (0.04)	−0.82 (0.04)	−0.83 (0.06)
R^2	0.78	0.73	0.79	0.74
Observations	8,689	146	120	77

Notes: The dependent variable is the urbanization rate.

the estimated intercepts and slopes are remarkably constant over time, which is reassuring when the relationship is used to predict agricultural employment shares out of sample. In the remainder of this section, I describe the methodology and data sources I use to accomplish this task.

4.1 Urbanization: Sources and definitions

Censuses not only provide information on the total size if the population, but also its location, thus allowing urbanization rates to be calculated. The United Nations has collected census

data on urbanization from all the member countries back to 1950.[14] Additional information from sample surveys and from estimates, made either by national governments or the United Nations, are used if no census data are available.[15]

The urbanization rates are not comparable across countries as the definition of urban varies. Most countries define urban areas in terms of the number of people living in a given agglomeration, but the cut-off ranges from just 200 in Denmark to 10,000 in Italy and Senegal. Some countries also require agglomerations to be primarily non-agricultural in terms of employment, to have access to electricity, or to be the administrative centre of a municipality before they are considered urban in the statistics.

The United Nations make no attempt to harmonize definitions, as they argue that a harmonization is unlikely to increase comparability.[16] A settlement of 5,000 people are often decidedly rural in terms of economic activity in China and India, whereas it is likely to be a centre of commerce and industry in Europe. Bairoch (1991) similarly argue that small towns in Europe historically have been peculiar in a global comparison due to the limited agricultural activity of their inhabitants. He offers Western Nigeria as an example of the opposite. In the 1952 census, roughly one third of the males living in cities with populations between 5,000 and 80,000 were engaged in agriculture. As it will be shown below, the estimation strategy I use is not affected by differences in definitions of urban.

Prior to 1950, I urban population data from Eggimann (1999) for Africa, Asia and Latin America, and from the USSR republics from Lewis *et al.* (1976). With a few exceptions, Countries in Western Europe, North America and Oceania have actual employment data back to the 19th century, and urbanization data are therefore not needed. Eggimann (1999) provides population estimates for named cities, not urbanization rates for entire countries. To calculate the

[14] United Nations (2012a).

[15] If no obervations are available in 1950, extrapolation is used to extend the data set backwards. For about 10 percent of the countries, the extrapolation is applied to a longer period than decade. Most of these countries are island nations with populations too small to be included in the data set of this paper (*e.g.*, Palau, Vanatu, and Tuvalu).

[16] United Nations (2012b).

latter, knowledge of total populations is needed. I obtain population data from International Historical Statistics augmented by various historical census data from other sources.[17]

The resulting urbanization rates are defined differently than the urbanization rates in the United Nations (2012a) data from 1950 and onwards. I use the overlapping observations in 1950 and Zipf's law to adjust the pre-1950 data to conform to the post-1950 definition (which differ from country to country).[18]

4.2 Estimation strategy

It is useful to illustrate where the empirical relationship between urbanization and agricultural employment comes from before the strategy for estimating agricultural employment out of sample is reviewed. The starting point is the following accounting equation for the labor force engaged in agriculture:

$$L_{i,t}^{A} = AES_{i,t}^{u} e_{i,t}^{u} U_{i,t} + AES_{i,t}^{r} e_{i,t}^{r} \left(P_{i,t} - U_{i,t} \right), \tag{1}$$

where $L_{i,t}^{A}$ is the number of individuals engaged in agriculture in country i in period t, $P_{i,t}$ is the total population, and $U_{i,t}$ is the number of individuals living in urban areas according to the national definition. $AES_{i,t}^{u}$ and $e_{i,t}^{u}$ are the agricultural employment share and the employment rate in urban areas. $AES_{i,t}^{r}$ and $e_{i,t}^{r}$ are the similar variables for rural areas.

Equation (1) can be rewritten to yield an expression for the aggregate agricultural employment share, $AES = \frac{L^{A}}{L}$:

[17] These include Lewis *et al.* (1976) for the USSR, McGee (1964) for Malaysia, Karpat (1985) for the Ottoman Empire, and McEvedy *et al.* (1978) for Indonesia and the countries on the Indian subcontinent. For Africa, I rely on population estimates made by Manning (2010), which correct known errors in the colonial records.

[18] Zipf (1941). The adjustments to the urbanization series are done as follows. Let u_x denote the urbanization rate where the cut-off for urban is cities with x inhabitants. Let u_y be defined in the same way. Zipf's law states that cities are distributed according to a power law. Power laws are scale invariant, and $\frac{u_x}{u_y}$ is consequently constant and independent of how the aggregate urban population evolves over time. It follows that $u_{x,t} = u_{y,t} \frac{u_{x,T}}{u_{y,T}}$, where t indexes time, and T is an arbitrary period where the two series overlap. This relationship allows $u_{x,t}$ to be calculated in periods where only $u_{y,t}$ is available, and vice versa.

$$AES_{i,t} = \left\{ \left(AES^u_{i,t} e^u_{i,t} - AES^r_{i,t} e^r_{i,t} \right) U_{i,t} + AES^r_{i,t} e^r_{i,t} P_{i,t} \right\} \frac{1}{L_{i,t}}.$$

Define the urbanization rate as $UR_{i,t} = \frac{U_{i,t}}{P_{i,t}}$, and let $e_{i,t}$ be the aggregate employment rate. It follows that:

$$AES_{i,t} = \left(AES^u_{i,t} \frac{e^u_{i,t}}{e_{i,t}} - AES^r_{i,t} \frac{e^r_{i,t}}{e_{i,t}} \right) UR_{i,t} + AES^r_{i,t} \frac{e^r_{i,t}}{e_{i,t}}.$$

This equation can be rearranged to yield the estimation equation used in Figure 2 and in Table 3:

$$UR_{i,t} = \beta_0 + \beta_1 AES_{i,t} + \varepsilon_{i,t} \tag{2}$$

Estimates of the two parameters β_0 and β_1 are are provided in Table 3 in the introduction to this section. The error term $\varepsilon_{i,t}$ contains idiosyncracies in agricultural employment shares and participation rates in rural and urban areas uncorrelated with the *aggregate* agricultural employment share. The idiosyncrasies are, in part, driven by differences in the national definitions of urban and rural populations. The agricultural employment share in urban areas will, for example, be larger when the towns defined as urban are smaller.

The uncorrelatedness of $\varepsilon_{i,t}$ is not sufficient to estimate $AES_{i,t}$ out of sample. Additional assumptions about the error term $\varepsilon_{i,t}$ are needed.

Growth rates in the agricultural employment share are proportional to growth in the urbanization rate if the error term $\varepsilon_{i,t}$ is assumed to be time invariant ($\varepsilon_{i,t} = \varepsilon_i$). The same is true for expected growth rates if $\varepsilon_{i,t}$ is i.i.d. However, both assumptions are unrealistic, since the error term $\varepsilon_{i,t}$ depends on employment rates and agricultural employment in urban and rural areas. They are behavioral variables that are likely to evolve over time, but only very gradually.

Instead, I assume that assume that the ratio $\frac{\varepsilon_{i,t}}{UR_{i,t}}$ is constant and equal to $\frac{\varepsilon_{i,T}}{UR_{i,T}}$ in periods $t < T$, where T is the earliest year with employment data, and thus the earliest year for which Equation (2) can be estimated. There are several reasons for choosing this specification. It is consistent with slow-moving behavioral variables, and the movement of the resiudal is toward

zero, which is reasonable if countries that are closer to an undeveloped steady state are more likely to be identical in terms of urbanization and agricultural employment than more developed countries. The variance of the error term should thus be lower when urbanization is lower. Furthermore, the assumption that $\frac{\varepsilon_{i,t}}{UR_{i,t}} = \frac{\varepsilon_{i,T}}{UR_{i,T}}$ for $t < T$ is essentially a way to impose mean reversion in the errors, since urbanization rates generally are increasing over time. Errors will have this property if there are measurement errors in the employment data in year T. Lastly, the assumption implies that $\gamma > 0$ in the following regression:

$$|\varepsilon_{i,t}| = \gamma UR_{i,t} + country\ fixed\ effects + error$$

That is indeed the case. $\gamma = 0.38$ with a t-value of 10.87.[19]

The assumption that $\frac{\varepsilon_{i,t}}{UR_{i,t}} = \frac{\varepsilon_{i,T}}{UR_{i,T}}$ is also practical when estimating the agricultural employment share out of sample. Substituting the assumption into Equation (2) yields:

$$\begin{aligned} UR_{i,t} &= \beta_0 + \beta_1 AES_{i,t} + UR_{i,t}\frac{\varepsilon_{i,T}}{UR_{i,T}} \\ &\iff UR_{i,t}\left(1 - \frac{\varepsilon_{i,T}}{UR_{i,T}}\right) = \beta_0 + \beta_1 AES_{i,t} \\ &\iff UR_{i,t}\frac{\widehat{UR}_{i,T}}{UR_{i,T}} = \beta_0 + \beta_1 AES_{i,t}\ , \end{aligned}$$

where $\widehat{UR}_{i,T}$ is the predicted value of the urbanization rate in period T obtained from the regression in Equation (2). The entire term $UR_{i,t}\frac{\widehat{UR}_{i,T}}{UR_{i,T}}$ corresponds to urbanization in period t, or $\widehat{UR}_{i,t}$.Differences in the national definition of urban are contained in the error terms of Equation (2), and the predicted urbanization rate can therefore be interpreted as being harmonized according to a definition that depends on the aggregate agricultural employment share.

The agricultural employment share in period t can be calculated as:

[19] Estimated for the full sample. The results are similar when the sample is limited to developing countries, and when the squared residuals are used as dependent variable in the regression.

$$AES_{i,t} = \frac{\widehat{UR}_{i,t} - \beta_0}{\beta_1},\ t < T \tag{3}$$

I use this equation along with the pooled estimates of the parameters, *i.e.*, $\beta_0 = 0.81$ and $\beta_1 = -0.82$, to obtain values of $AES_{i,t}$ in years prior to the year where actual employment data become available. Some countries have not yet published data on agricultural employment shares in the last few years of the sample period, so a similar approach is used to estimate agricultural employment shares in these years.

4.3 Precision

To test the accuracy of the methodology to estimate AES, I set the starting year $T = 2000$ for all countries. I then derive the predicted agricultural employment shares for 1970 and 1950, and compare the predictions to the observed agricultural employment shares by estimating the following regressions:

$$AES_{i,t} = \lambda_1 \widehat{AES}_{i,t} + \eta_{i,t}, \tag{4}$$

and

$$AES_{i,2000} - AES_{i,t} = \lambda_2 \left(\widehat{AES}_{i,2000} - \widehat{AES}_{i,t} \right) + \mu_{i,t}, \tag{5}$$

for $t = 1950, 1970$. The estimated parameters should be $\lambda_1 = \lambda_2 = 1$ if the predicted agricultural employment shares are unbiased, and the prediction errors, $\eta_{i,t}$ and $\mu_{i,t}$, should be small if the estimates are accurate. The results are shown in Table 4. For $t = 1950$, both estimated parameters are somewhat smaller than one, but not significantly so at a 95 percent confidence level. For $t = 1970$, λ_1 and λ_2 are are significantly smaller than one in a statistical sense, but the difference does not seem substantial in the case of λ_1.

The estimate of λ_2 is 0.85, indicating that the estimated changes in agricultural employment shares may overstate the actual changes. An explanation is that idiosyncratic shocks play a

Table 4: Test of AES estimation

Year	**1950**	**1950**	**1970**	**1970**
Estimated parameter	λ_1	λ_2	λ_1	λ_2
Estimate	0.98 (0.02)	0.95 (0.03)	0.95 (0.02)	0.85 (0.04)
R^2	0.96	0.90	0.96	0.78
Observations	77	77	114	114

Notes: Results of estimating Equation (4) and (5)

larger role when the time frame is short, and the estimated λ_2 is therefore harder hit by attenuation bias. Consistent with this interpretation, the estimated λ_2 falls further when the time period is shortened, and the standard error increases despite more observations. For $t = 1980$, for instance, $\lambda_2 = 0.75$ with a standard error of 0.06. A small bias in the short run of this sort is a minor worry, as the aim of the database is to provide a measure of long run growth. Moreover, the very high R^2 indicate that the prediction errors are relatively minor. Based on the results in Table 4, I therefore consider the estimated agricultural employment shares to be accurate.

5 The final data set

Three additional steps are taken to increase the coverage of the data set. First, I interpolate between observations. Census data are, for instance, usually collected once every decade, and the census year differ from country to country. I use a simple linear interpolation to fill the gaps if no other data source is available. The second step is to place an upper bound on the estimated agricultural employment shares. The highest agricultural employment share observed in a census or a survey is 0.94 (Nepal 1971). It seems unlikely that any country have exceeded that number by much in modern times, since traders, craftsmen and government administrators exist in even the most underdeveloped nations. I therefore set $AES_{i,t} = 0.95$ if the predicted values from Equation (3) are bigger than 0.95.[20] The third step is to extrapolate

[20] A similar boundary problem can potentially arise when the predicted $AES_{i,t}$ is close to 0. However, all countries with low employment shares in agriculture are developed at the time of measurement, and have survey

the agricultural employment shares backward in the eight countries where the upper bound of 0.95 is reached in the first year of observation.[21] The implicit assumption is that this group of countries are in an underdeveloped steady state. To confirm this assumption, I cross check with the urban population data in Eggimann (1999). None of the countries for which the extrapolation is made had any significant urban development in 1900.

The coverage of the final database is illustrated in Figure 3. From 1950 an onward, the database contains information on agricultural employment shares for 169 countries out of the 178 countries in the world with more than 250,000 inhabitants in 2010. The number falls to 116 countries in the beginning of the 20th century. Most of the recent data points are based on actual employment data, whereas estimates based on urbanization rates are more prominent in the beginning of the period. All countries have at least one observation based on actual employment data.

There are about 30 countries with no data prior to 1950. They are spread evenly across the developing world. The exception is the countries on the Arab Peninsula, where no census or survey data were collected before World War II.

To track the long run growth process for more countries than what is possible in other data sets currently available to researchers is a central motivation for constructing the database. The success criterion is therefore two-fold. The database should have a broader coverage than existing data sets, and it should be an accurate yard stick of development.

Success on the first criterion can easily be judged by comparing the data availability to other data sets containing information on national income or development. By construction, my database covers more countries in more years than any of the individual sources of employment data it is based on. One of the most well-known and comprehensive data sets for analyzing national income over the long run is Angus Maddison's historical GDP data.[22] The black line

or census data on AES. No estimation is therefore needed.

[21] The countries in question are Bhutan, Burundi, Lesotho, Nepal, Niger, Papua New Guinea, Rwanda and Zambia.

[22] Maddison (2010).

Figure 3: Data availability

in Figure 3 shows the number of countries for which Maddison provide GDP estimates.[23] It is below the number of countries in my database for all years, and my database thus compares favorably with Maddison in terms of coverage.

The comparison to Maddison's GDP data is, obviously, only relevant if the employment data are useful for analyzing income levels or economic growth. The next section evaluates my data set along this dimension.

6 The evolution of agricultural employment and GDP

The share of the labor force employed in agriculture in a country today is, as argued in Section 2, a an accurate predictor of its national income. In this section, I show that the same has been true historically, and, by implication, that agricultural employment shares are suitable to analyze long run trends in income. I also demonstrate that agricultural employment shares are

[23] There are gaps in the Maddison GDP data for many countries. I have counted years where interpolation makes it possible to fill gaps as observed when assessing the data availability in Maddison.

Table 5: Agricultural employment and historical GDP estimates

	1900	**1925**	**1950**	**1975**	**2000**	**2000***
Constant	9.06	8.86	9.23	9.55	9.51	9.18
	(0.14)	(0.12)	(0.12)	(0.08)	(0.07)	(0.11)
Coefficient on *AES*	−2.82	−2.39	−3.06	−3.36	−3.97	−3.38
	(0.21)	(0.19)	(0.17)	(0.14)	(0.18)	(0.18)
R^2	0.80	0.72	0.71	0.79	0.76	0.67
Observations	48	66	134	157	158	120

*Notes: *Countries with AES < 0.1 are excluded. The dependent variable is log GDP per capita from Maddison (2010).*

able to correctly identify well-known measurement errors in the GDP data. My database can therefore be a useful alternative to GDP data when studying economic growth and comparative development, even in periods when GDP data are available.

6.1 Stability

To check whether the link between agricultural employment and income has been stable over time, I regress Maddison (2010) log GDP per capita on the agricultural employment shares from my database. The results are reported in Table 5 for different years. The constant term seems to be increasing slightly over time, and the coefficient on the agricultural employment share falls from -2.82 in 1900 to -3.97 in 2000.

Part of the explanation is that the sample changes over time. Data become available for more and more countries, and the sample of countries becomes more representative in terms of the stages of development. While the 1900 sample includes all the most developed nations at the time, only the United Kingdom had an agricultural employment share below one quarter. In 2000, on the other hand, 86 countries had an agricultural employment share below one quarter. The difference is illustrated in Figure 4 by the data points and regression lines for the years 1900 and 2000.

The reason why the slope and intercept of the regression change when the sample includes more highly developed countries is that countries continue to grow after the agricultural employment share has fallen to near zero. That is the case for the cloud of observations above

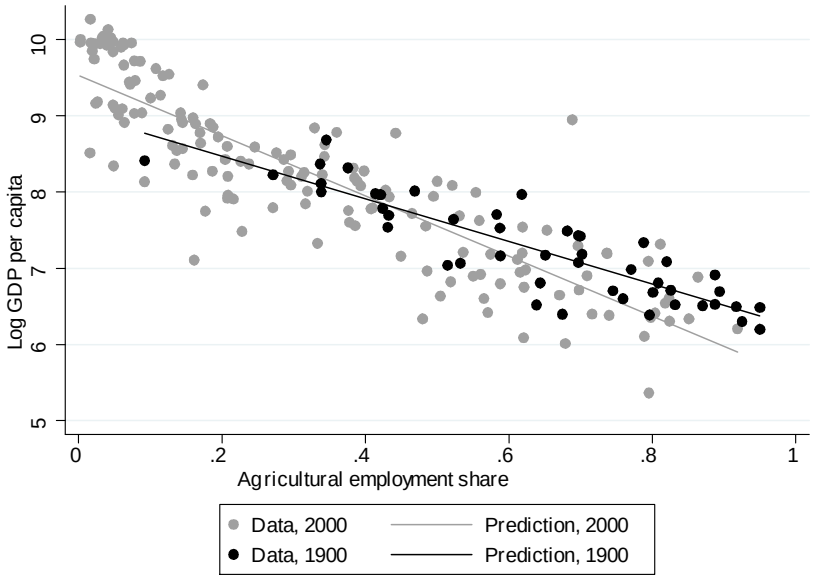

Figure 4: Agricultural employment and real GDP per capita.

the regression line in the upper left corner of Figure 4. To demonstrate the consequence of the regression estimates, countries with less than ten percent of the workforce employed in agriculture in 2000 are excluded from the sample in the final column of Table 5. Compared to the full sample, the intercept is lower and the slope flatter. Moreover, the regression line is no longer significantly different from the one estimated for 1900. The relationship between Maddison's GDP estimates and the agricultural employment shares therefore seems to have been fairly stable over time in countries where agriculture still plays a significant role in the economy.

6.2 Regional GDP growth

Figure 5 and 6 show the evolution of GDP per capita and agricultural employment in the 20th century for eight regions. In each region, the two variables are population weighted. Countries with missing observations in parts of the period are left out, but the subsamples are nonetheless fairly representative for the region. The exception is North Africa and the Middle East grouping, which does not cover the Arab Peninsula. The scales for agricultural

employment are inverted, and adjusted such that a given value of the agricultural employment share on the left hand y-axis corresponds to the predicted GDP level according to the regression for 1950 reported in Table 5.

Panel A of Figure 5 shows that GDP per capita in Western Europe closely resembles income as predicted by agricultural employment from 1900 to the 1960s. From then, GDP has outpaced the decline in the agricultural employment share. The same pattern is found in the European off-shoots (Australia, Canada, New Zealand and the United States), where the growth rates in the two variables also decoupled in the 1960s. The decoupling of GDP from agricultural employment is a sign that the economies have completed the transition out of agriculture.

Although the trend is as expected, the level of GDP per capita in the European off-shoots prior to 1960 is higher than predicted by their agricultural employment shares. The low population densities and the vast expanses of fertile land presumably account for this finding. The high land-to-labor ratios made agriculture, at a given stage of development, more attractive in the United States and in the other off-shoots than in Western Europe.

Before the fall of communism, GDP was higher than indicated by agricultural employment in the USSR and in Eastern Europe. Like the European off-shoots, USSR had a low population density, which could explain the observed pattern. Other explanations are possible, however. Output, as measured by GDP, declined markedly after the fall of communism. But electricity consumption declined much less than what should be expected given the size of the contraction, indicating that the output decline probably is exaggerated in the GDP statistics.[24] One source of inflated GDP numbers is overcounting of produced quantities, which were common in the communist era when factory managers had to reach certain production targets every year.[25] Another source of inflated GDP numbers is the price data. The heavy industries in the USSR and Eastern Europe produced large quantities of low quality goods that would have been in limited demand in a market economy. The official prices of these goods before the fall of

[24] Eichengreen (2008).

[25] Åslund (2001).

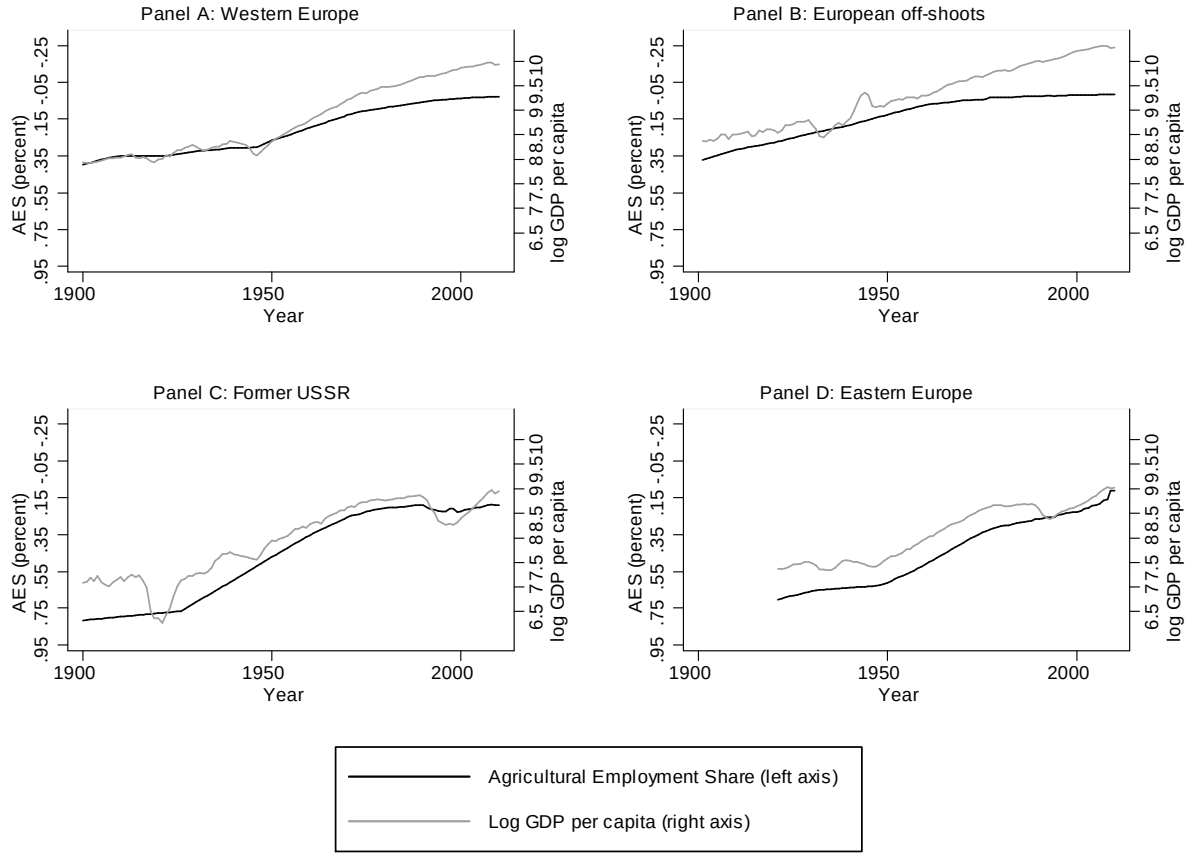

Figure 5: Long term growth in GDP/capita and the agricultural employment share (inverted scale)

communism are likely to have exaggerated their values. And observed market prices *after* the fall of communism, when supply was reduced and quality increased, will similarly be too high. Using either of these alternatives cause the GDP statistics overstate income levels before the end of communism, but no other price data are available to national accountants. The comparison with agricultural employment shares in Figure 5 makes the overestimation of GDP for USSR and communist Eastern Europe clearly visible.

Figure 6 shows agricultural employment shares and GDP per capita in four regions that largely consist of low and middle income countries. Japan, Singapore and South Korea are obvious exceptions, but these countries have a relatively small impact on Asia as a whole, given

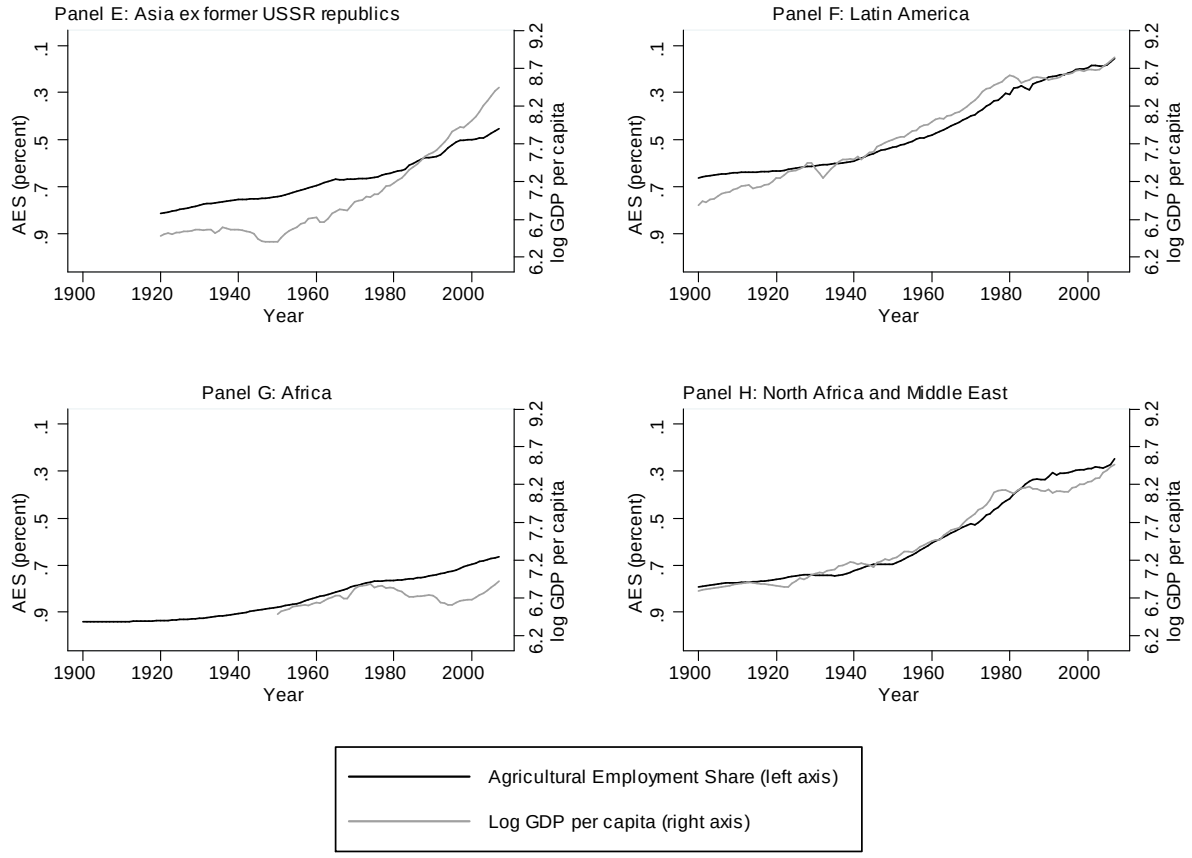

Figure 6: Long term growth in GDP/capita and the agricultural employment share (inverted scale)

the large populations in China, Indonesia and on the Indian subcontinent. The scales on the y-axes are adjusted slightly from the ones in Figure 5 to make the trends in the data more visible.

GDP per capita has largely evolved as predicted by the agricultural employment share in Latin America, and in North Africa and the Middle East. The agricultural employment share in Asia, on the other hand, indicates that incomes were substantially higher in the middle of the 20th century than what is implied by GDP estimates, and that the subsequent growth miracle therefore was less dramatic. Explanation this finding is an interesting topic for further research.

6.3 Income in sub-Saharan Africa

The GDP estimates in sub-Saharan Africa also conflict with the observed fall in the agricultural employment share. As it can be seen in Panel G of Figure 6, the two series follow each other closely from 1950 to the 1970s. In the following to decades, GDP fell, and a sizeable gap to the income level predicted by agricultural employment and the GDP estimates emerged. Research by Young (2012) and Jerven (2013) show that GDP estimates in Africa suffer from a significant downward bias. To demonstrate that the bias account for most of the gap observed in Figure 6, I compare the income level for sub-Saharan Africa implied by agricultural employment shares to the level predicted by another proxy for income: Night lights measured from space by sattelites.

To do so, I first regress measured log GDP per capita on agricultural employment and a dummy variable for sub-Saharan Africa for the year 2000. As in the final column of Table 5, I exclude countries with less than 10 percent of the labor force employed in agriculture from the sample. The results are reported in Table 6. The coefficient on the dummy variable indicates that the measured GDP level in sub-Saharan Africa is approximately 28 percent lower than the income level predicted by agricultural employment shares.

The last column of Table 6 shows the results of a similar regression where light intensity per capita rather than agricultural employment is used as a proxy for true income. Consistent with the agricultural employment shares, the night lights imply that GDP in sub-Saharan Africa is underestimated by 25 percent.

6.4 Convergence and global inequality

The slightly different growth paths of agricultural employment shares and GDP per capita have implications for the measured income dispersion between countries. The patterns of global inequality depend on which of the two proxies for income that is used to compute them. Figure 7 shows how the standard deviations across countries of agricultural employment shares and of log GDP per capita have evolved over time. A decline in the standard deviation is known

Table 6: GDP bias in Africa

Explanatory variable	*AES*	Lights/capita
Constant	9.15 (0.11)	10.15 (0.16)
Coefficient on explanatory variable	−3.02 (0.27)	0.48 (0.04)
Coefficient on Africa dummy	−0.28 (0.11)	−0.25 (0.12)
R^2	0.69	0.73
Observations	115	118

Notes: The dependent variable is log GDP/capita from Maddison (2010). The light data are from Henderson et al. *(2012)*

as σ-convergence.

Panel A shows convergence for the 49 countries with data available both in Maddison (2010) and in my database since 1900. The standard deviation of agricultural employment shares is rescaled to make it directly comparable to the standard deviation of log GDP per capita.[26] Both the Maddison GDP data and the agricultural employment shares imply a stable degree of global inequality in the first half of the 20th century. From then there has been σ-divergence, according to the GDP numbers, and global inequality was at an all-time high around the end of the century. By contrast, the employment data show σ-convergence.

A similar conclusion is reached when using GDP data from Penn World Tables 8.0 (Panel B, 104 countries), whereas the employment data and GDP data from the Barro-Ursua macroeconomic data set, used to study convergence by Barro (2012), are more in agreement (Panel C, 34 countries).

The difference between the two series in Panel A and B is partly driven by the downward

[26] I use the regression coefficient α_1 from the following regression to scale the standard deviations of agricultural employment shares: $\log(GDP/capita) = \alpha_0 + \alpha_1 AES + error$. The regression corresponds to the ones reported in Table (5), but is estimated for the pooled sample of all years. The implicit assumption is that α_1 is stable over time. The results of Table (5) suggest that the assumption is a reasonable approximation in countries where agricultural employment is still significant. Admittedly, the point estimate of α_1 has been rising (numerically) over the 20th century. While the rise is not significant in a statistical sence, it may, if it reflects an actual change in the structural relationship, cause a slight secular tendency to σ-convergence in the agricultural employment data. This potential bias strengthen the results presented in this sub-section.

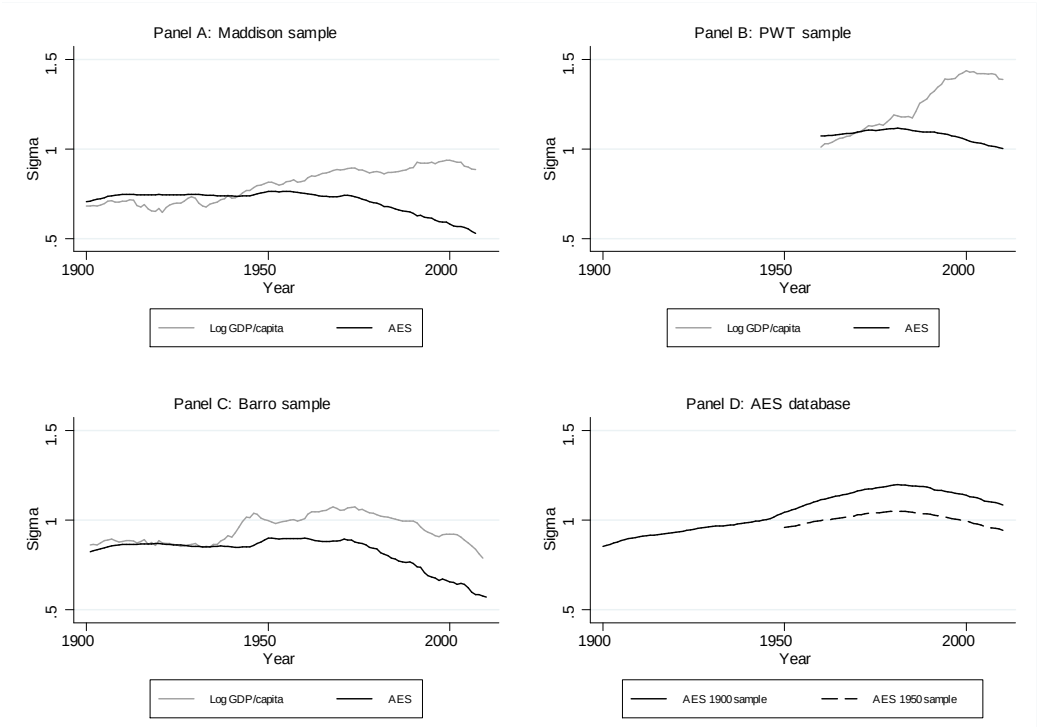

Figure 7: Sigma convergence

bias in GDP data for Africa and other developing countries. The most visible sign of this is the marked σ-divergence in the Penn World Tables data relative to the σ-divergence in the Maddison data, which contain fewer countries from the developing world.

The σ-convergence in agricultural employment shares is also a consequence of the variable being bounded between 0 and 1. By construction, global inequality in agricultural employment shares is also bounded. Such boundedness naturally gives rise to what amounts to a global Kuznets curve for structural transformation. Inequality was low in the premodern era, where all countries had most of their labor forces employed in agriculture, and will similarly be low when all countries have completed the transition out of agriculture. Between these two extremes, differentiated timing of the transitions out of agriculture imply high degrees of inequality.

The Kuznets curve is clearly visible in Panel D of Figure 7, where the σ for the full sample of countries available in my database is shown. The 1900 sample consists of the 116 countries

with data for the full period, and the 1950 sample is the 169 countries with data available from 1950 and onward. It is interesting to note that peak inequality coincide with the onset of a new era of globalization.

The presence of a Kuznets curve in agricultural employment shares shows that initially rich countries, *i.e.*, countries that had begun the the transition out of agriculture, grew faster than other countries in the first half of the 20th century. This is consistent with Unified Growth Theory, where initial conditions explain differentiated timing in the transition to modern economic growth.[27] It is, however, at odds with the "iron law" of conditional β-convergence, which states that poor countries will have faster GDP per capita growth than rich ones.

Estimating rates of conditional β-convergence is suprisingly difficult, not just because of measurement errors in GDP. The econometrics is tricky, and a consensus on the best empirical strategy has yet to emerge.[28] Estimates based on GDP data face the additional problem of limited numbers of observations. Researchers can either estimate the convergence rates from the large and fairly representative sample of countries with GDP data available from the 1960s or 1970s ("large N, small T"), or from the much more restricted sample of countries with data going back to the turn of the 20th century ("small N, large T"). Estimated convergence rates in the large-N-small-T sample will not pick up possible β-divergence in the first two thirds of the 20th century. Estimated convergence rates using a longer small-N-large-T sample are, on the other hand, unlikely to be representative. For example, Barro (2012) finds an upper bound of the rate of convergence of 2.4 percent since 1870 in the sample of 34 countries also shown here in Figure 7, panel C.[29] Inequality, as measured by dispersion in agricultural employment shares, was stable or declining (σ-convergence) among the 34 countries analyzed by Barro in the entire period. By contrast, inequality was rising until the 1970s in the larger sample of 116 countries in panel D, and they are more unequal today than in 1900. σ-divergence do not necessarily

[27] *E.g.*, Galor (2005, 2011).

[28] See Durlauf *et al.* (2005) for an overview of the debate.

[29] It is an upper bound, as a finite time period and the presence of fixed effects in the regression causes an upward Nickell (1981) bias in the estimate.

imply conditional β-divergence. Still, the rising inequality in the larger sample indicates that the rate of β-convergence estimated by Barro (2012) and others in small-N-large-T samples may be biased upward due to selection.

My agricultural employment share database provide a sample that is large in both the N and T dimensions. It is an interesting topic for further research to see if the results of the convergence literature, and the empirical growth literature more broadly, are robust to using the larger sample of agricultural employment shares rather than GDP data.

6.5 Predicting GDP revisions with agricultural employment

As shown above, well-known measurement errors in GDP in Africa, USSR and Eastern Europe are made clearly visible by comparing GDP data to agricultural employment shares. But agricultural employment shares are in general useful for detecting and correcting measurement errors in GDP data. To illustrate this, I use my database to predict GDP revisions following the publication of new benchmark estimates for PPPs. Predicting revisions is equivalent to predicting measurement errors if the revisions, on average, make GDP estimates more accurate. The excercise is related to Almås (2012), who use consumption data and Engel's Law to correct PPP estimates.[30]

The benchmark year for PPPs is changed when the results of a new ICP round is published. Updates happen infrequently, and often have huge impacts on relative GDP levels. Chinese real GDP was, for instance, reduced by 39 percent in 2008 following the publication of the results from the 2005 ICP. China's place in the world economy made popular media pay much attention to this change, but the size of the revision was not unique to the Middle Kingdom. As shown in Table 7, 24 countries had their estimates of real GDP per capita revised by more than China. The benchmark PPPs were revised once more when the results of the 2011 ICP were published in 2014, and the revisions were of the same magnitude as in the 2005 round.

[30] Hamilton (2001) and Costa (2001) similarly use Engel's Law to correct consumer price indecies in the United States, with implications for estimates of real GDP growth.

Table 7: Revisions to PPP adjusted GDP after the 2005 ICP round

Country	Revision (%)	Country	Revision (%)
Congo	188	Ghana	-51
Yemen	137	Angola	51
Gabon	94	Ecuador	50
Lebanon	84	Comoros	-47
Zimbabwe	-74	Cambodia	-47
Nigeria	73	Venezuela	47
Kuwait	71	Ethiopia	-46
Congo, Dem. Rep.	-63	Central African Rep.	-45
Gambia	-62	Tanzania	44
Guinea	-60	Philippines	-43
Lesotho	-58	Namibia	-40
Cabo Verde	-51	Togo	-40

Source: The World Bank (2008), Appendix G

To predict the GDP revisions, I use a two step procedure. In the first step, I estimate the following regression (corresponding to the regression lines in Figure 1 and Figure 4):

$$\log\left(\widetilde{y}_i^{old}\right) = \alpha_0 + \alpha_1 \widetilde{AES}_i + \varepsilon_i^{old}, \tag{6}$$

where $\widetilde{y}_i^{old}$ is measured GDP per capita before the revision, $\widetilde{AES}_i$ is the measured agricultural employment share, and ε_i is an error term. The latter can be decomposed into three parts:

$$\varepsilon_i^{old} = \varepsilon_i^{y,old} - \alpha_1 \varepsilon_i^{AES} + \varepsilon_i^{model}. \tag{7}$$

The first term on the right hand side, ε_i^y, is the measurement error in GDP per capita defined such that the true unobserved income level is given by $\log(y_i) = \log\left(\widetilde{y}_i^{old}\right) - \varepsilon_i^{y,old}$. The estimated residuals are thus correlated with the measurement error in the unrevised GDP numbers, and can be used to correct the GDP estimates by following the approach of Henderson *et al.* (2012), who use a weighted average of observed GDP and night lights measured from space as a measure of true GDP. The correlation is also useful to predict revisions to GDP. A revision is given by:

$$\begin{aligned}\Delta\varepsilon_i^y &= \log\left(\widetilde{y}_i^{new}\right) - \log\left(\widetilde{y}_i^{old}\right) \\ &= \varepsilon_i^{y,old} - \varepsilon_i^{y,new} \\ &= \varepsilon_i^{old} - \alpha_1\varepsilon_i^{AES} + \varepsilon_i^{model} - \varepsilon_i^{y,new}\end{aligned} \tag{8}$$

The terms $\varepsilon_i^{y,new}$, $\alpha_1\varepsilon_i^{AES}$ and ε_i^{model} are unobserved, but ε_i^{old} can be obtained from estimating Equation (6). If the revisions improve the income estimates, meaning that $cov\left(\Delta\varepsilon_i^y, \varepsilon_i^{y,old}\right) < 0$, then the parameter β_1 should be positive in a second step, where the revisions are regressed on the residuals from Equation (6):

$$\Delta\varepsilon_i^y = \beta_0 + \beta_1\varepsilon_i^{old} + u_i \tag{9}$$

The results of the first step are shown in Table 8, Panel A, for both the 2005 and the 2011 ICP rounds. Results from the second step, *i.e.*, the regression in Equation (9), are reported in Panel B. The estimates of β_1 are significant with $p < 0.01$ in both years, confirming that agricultural employment data are useful to predict revisions to GDP, and, by implication, useful for detecting measurement errors in the GDP data.

The R^2 is not particularly high in any of the two regressions in Panel B of Table 8 for three reasons: Country specific idiosyncrasies affects the relationship between the true income level and agricultural employment (*e.g.*, high land-to-labor ratios), measurement errors introduce noise in the agricultural employment data, and the GDP revisions did not remove all of the noise in the GDP data. The latter is especially important due to the many sources of measurement errors in GDP discussed in Section 2. It is also underscored by the need for consecutive large revisions to the PPPs.

The large revisions to the PPPs, and the possibility of predicting them using agricultural employment shares, emphasize that agricultural employment is a useful proxy for income. Moreover, both the PPP data, and the other data underlying estimates of real GDP, are likely to be less accurate further back in time than in the periods analyzed here. The usefulness of

Table 8: Predicting PPP revisions

Panel A: Results of regression (6)		
	2005	**2011**
$\hat{\alpha}_1$ (estimated coefficient on AES_i)	-3.99	-4.43
Standard error	0.21	0.21
R^2	0.75	0.76
Observations	130	141

Panel B: Results of regression (9)		
	2005	**2011**
$\hat{\beta}_1$ (estimated coefficient on ε_i^{old})	0.14	0.08
Standard error	0.05	0.02
R^2	0.06	0.07
Observations	130	141

Notes: The unrevised 2011 GDP estimates are calculated by the author using the constant PPP approach.
Sources: The World Bank (2008, 2014) and the United Nations National Accounts Main Aggregates Database.

agricultural employment as an alternative to GDP data is thus likely to increase further back in time.

7 Concluding remarks and avenues for further research

I have collected and estimated agricultural employment shares for 169 countries in the period 1900-2010. Agricultural employment is closely related to national income, and reliable data are available for more countries and for longer periods than GDP estimates. The resulting database, available online, is useful for researchers studying economic growth, comparative development, economic history, or other related fields.

An obvious application of the database is to test the many theories directly concerned with agricultural employment, such as models of structural change, theories of how industrialization spread around the globe, and dual economy models. Other potential applications have been noted in this paper. Agricultural employment shares are, for instance, useful to detect and

correct measurement errors in GDP. Perhaps more importantly, the availability of long time series for employment makes it possible to study long run growth and convergence in more representative samples of countries than the ones usually analyzed in the literature.

The larger sample of countries is likely to matter for the conclusions drawn. Inequality, measured by the dispersion of agricultural employment shares, was stable in the first three quarters of the 20th century among the countries with long time series of GDP data available in Maddison (2010) and in the Barro-Ursua data set. But among the countries in my larger sample, inequality rose by 50 percent during that period, corresponding to σ-divergence. This finding suggests that existing estimates of β-convergence rates may be overestimated as well. Estimating rates of β-convergence based on agricultural employment is an interesting topic for further research. And more generally, it will be interesting to see if the results of the empirical growth literature are robust to replacing GDP per capita with agricultural employment shares as the variable of interest.

Agricultural employment are also linked to many other important outcomes. The transition out of agriculture may, for instance, reduce fertility as it is harder to raise children when working outside of the home. It may also change the social standing of women, as nonagricultural work often is less physically demanding. Trade policies may likewise be affected when the composition of output changes. These are just a few examples. Changed sectoral employment patterns are likely to correlate with many other demographic, economic, social, and institutional variables.

It will also be interesting to study how agricultural employment is affected by global macro trends, such as the introduction of new technologies, lower transportation costs, and climate change. Understanding how climate change have interacted with agriculture around the world historically is particularly important, as it will give an indication of the economic consequences of further emissions of greenhouse gases.

Data collection projects, as the one presented here, are never entirely completed. It is my intention to update the database if, or when, more data come to light, and it is my hope that users of the database will contribute to this task.

Data Appendix

This appendix gives an overview of the database and its underlying sources. More details can be found in the additional data in Wingender (2014).

All independent countries with more than 250,000 inhabitants in 2010 are listed in the tables on the next pages. The second column of the tables contains the first year with available data. The third column contains the first year where the data point is based on employment data rather than estimated from urbanization rates. The data sources used are listed in the final column.

The following abbreviations are used in the final column for sources of employment data:

- ILO: International Labor Organization
- GGDC: Groningen Growth and Development Centre
- OECD: Organization for Economic Cooperation and Development
- IHS: International Historical Statistics from Mitchell (1993, 1998a,b).
- EF: Easterly and Fischer (1995)
- OLA: Oxford Latin America Economic History Database
- LoN: Statistical Yearbook of the League of Nations
- IPUMS: Calculations based on intregrated public use micro data from Sobek *et al.* (2013)
- NAPP: Calculation based on micro data from the North Atlantic Population Project by Minnesota Population Center (2008)
- My: Myers and Campbell (1954)
- NSO: Data collected from national statistics offices
- Good: Census data from the Habsburg empire kindly provided by David Good. The data are used in Good (1994) and Good and Ma (1998).

The following abbreviations are used in the final column for sources used to calculate urbanization rates:

- UN: United Nations
- IHS: as above
- Eg: Eggimann (1999)
- Ma: Manning (2010)
- Le: Lewis *et al.* (1976)

- Ka: Karpat (1985)
- Mc: McEvedy *et al.* (1978)

The remaining data sources are not abbreviated.

Country	First obs.	First AES obs.	Sources
Afghanistan	1920	1979	IHS, UN, Eg
Albania	1930	2002	ILO, NSO, UN, IHS
Algeria	1900	1948	ILO, IHS, UN, Eg
Angola	1900	1960	NSO, IHS, Eg, Ma
Argentina	1900	1902	GGDC, OLA, Eg, IHS
Armenia	1900	1970	ILO, EF, Le
Australia	1901	1901	IHS, ILO, OECD, GGDC
Austria	1900	1900	IHS, ILO, OECD, Good
Azerbaijan	1900	1970	ILO, EF, Le
Bahamas	1950	1973	ILO, UN, IHS, Eg
Bahrain	1950	1979	ILO, UN
Bangladesh	1900	1951	ILO, IHS, Eg, Mc
Barbados	1946	1946	ILO, IHS
Belarus	1900	1970	ILO, EF, Le
Belgium	1900	1900	OECD, GGDC, IHS
Belize	1950	1993	ILO, UN
Benin	1900	2003	ILO, UN, Eg, Ma
Bhutan	1900	2003	ILO, UN
Bolivia, Plurinational State of	1900	1950	GGDC, ILO, Eg, IHS
Bosnia and Herzegovina	1900	1900	ILO, IHS, My, Almanach de Gotha (1910)
Botswana	1950	1964	GGDC, UN
Brazil	1900	1900	GGDC, IHS, OLA, IPUMS
Brunei Darussalam	1950	1991	ILO, NSO, UN
Bulgaria	1900	1910	ILO, IHS, LoN, Lampe (1975)
Burkina Faso	1900	1985	ILO, IPUMS, UN, Eg, Ma
Burundi	1900	1979	ILO, IHS, UN, Eg, Ma
Cabo Verde	1950	2000	NSO, UN
Cambodia	1920	1962	ILO, IHS, UN, Eg,
Cameroon	1900	1976	ILO, IHS, UN, Eg, Ma
Canada	1900	1900	ILO, IHS, NAPP
Central African Republic	1900	1975	IHS, UN, Eg, Ma
Chad	1900	1993	ILO, NSO, UN, Eg, Ma
Chile	1900	1907	ILO, OLA, GGDC, Eg, IHS
China	1900	1980	ILO, IHS, Eg
Colombia	1900	1938	GGDC, ILO, OLA, Eg, IHS
Comoros	n.a.	n.a.	
Congo	1900	2005	ILO, UN, Eg, Ma
Congo, the Democratic Republic of the	1900	1952	IHS, UN, Eg, Ma
Costa Rica	1900	1950	GGDC, ILO, Eg, IHS
Côte d'Ivoire	1900	1964	NSO, IHS, UN, Eg, Ma
Croatia	1900	1900	ILO, IHS, My, Good
Cuba	1900	1919	OLA, ILO, IHS, Eg,

Country	First obs.	First AES obs.	Sources
Cyprus	1946	1946	KLEM, ILO, IHS
Czech Republic	1900	1900	OECD, GGDC, IHS, Good
Denmark	1900	1900	OECD, GGDC, IHS
Djibouti	n.a.	n.a.	
Dominican Republic	1900	1920	OLA, ILO, IHS, Eg
Ecuador	1900	1950	OLA, ILO, IHS, Eg
Egypt	1900	1907	ILO, IHS, Eg
El Salvador	1920	1950	OLA, ILO, IHS, Eg
Equatorial Guinea	1950	1983	ILO, UN
Eritrea	n.a.	n.a.	
Estonia	1900	1922	OECD, GGDC, ILO, EF, Lon, Le
Ethiopia	1900	1971	GGDC, IHS, UN, Eg, Ma
Fiji	1950	1956	IPUMS, IHS, UN
Finland	1900	1900	OECD, GGDC, IHS
France	1900	1900	OECD, GGDC, IHS
Gabon	1950	1963	ILO, IHS, Deldycke *et al.* (1968), UN
Gambia, The	1900	1993	ILO, UN, Eg, Ma
Georgia	1900	1970	ILO, EF, Le
Germany	1900	1900	OECD, IHS
Ghana	1900	1960	GGDC, IHS, UN, Eg, Ma
Greece	1920	1920	OECD, GGDC, IHS
Guatemala	1900	1950	OLA, ILO, IHS, Eg
Guinea	1950	1983	ILO, IPUMS, UN
Guinea-Bissau	n.a.	n.a.	
Guyana	1900	1946	ILO, IHS, Eg
Haiti	1900	1950	OLA, ILO, IPUMS, IHS, Eg
Honduras	1900	1950	OLA, ILO, IHS, Eg
Hong Kong	1920	1974	ILO; GGDC, UN, Eg, IHS
Hungary	1900	1900	OECD, ILO, IHS, Good
Iceland	1900	1900	ILO, LoN, NAPP
India	1900	1901	GGDC, ILO, Eg, Mc
Indonesia	1900	1905	GGDC, ILO, IHS, Eg, Mc
Iran, Islamic Republic of	1900	1956	ILO, IHS, UN, Eg
Iraq	1900	1957	ILO, IHS, UN, Eg
Ireland	1900	1900	OECD, GGDC, IHS, LoN
Israel	1948	1948	OECD, ILO, IHS
Italy	1900	1900	OECD, GGDC, IHS
Jamaica	1920	1943	ILO, IHS, Eg
Japan	1900	1872	OECD, GGDC, IHS
Jordan	1950	1961	ILO, IHS, UN
Kazakhstan	1900	1970	ILO, EF, Le

Country	First obs.	First AES obs.	Sources
Kenya	1900	1969	GGDC, UN, Eg, Ma
Korea, Democratic People's Republic of	n.a.	n.a.	
Korea, Republic of	1955	1955	OECD, GGDC, IHS, Chung (2006)
Kosovo	1921	1921	NSO, IHS, My
Kuwait	1950	1983	ILO, UN
Kyrgyzstan	1897	1970	ILO, EF, Le
Lao People's Dem. Rep.	1950	1995	NSO, UN
Latvia	1900	1925	GGDC, ILO, EF, LoN, Le
Lebanon	1910	1970	NSO, IHS, UN, Eg, Ka
Lesotho	1900	1999	NSO, ILO, UN, Eg, Ma
Liberia	1900	1962	ILO, IHS, UN, Eg, Ma
Libya	1900	1964	ILO, IHS, UN, Eg, Ma
Lithuania	1900	1923	GGDC, ILO, EF, LoN, Le
Luxembourg	1907	1907	OECD, GGDC, Deldycke *et al.* (1968)
Macedonia, the FYD of	1921	1921	ILO, IHS, My
Madagascar	1900	1993	NSO, ILO, UN, Eg, Ma
Malawi	1900	1966	GGDC, UN, Eg, Ma
Malaysia	1920	1947	GGDC, ILO, IHS, Eg, McGee (1964)
Maldives	1950	1990	NSO, ILO, UN, Eg, Ma
Mali	1900	1976	ILO, IHS, UN, Eg, Ma
Malta	1948	1948	KLEMS, ILO, Deldycke *et al.* (1968)
Martinique	1950	1961	ILO, IHS, UN
Mauritania	n.a.	n.a.	
Mauritius	1900	1952	GGDC, IHS, UN, Eg, Ma
Mexico	1900	1910	OECD, GGDC, OLA, Eg, IHS
Moldova, Rep. of	1900	1970	ILO, EF, Le
Mongolia	1920	1993	ILO, UN, Eg, IHS
Montenegro	1921	1921	ILO, IHS, My
Morocco	1900	1952	ILO, IHS, UN, Eg, Ma
Mozambique	1900	1950	ILO, IHS, Eg, Ma
Myanmar	1920	1978	ILO, Eg, IHS
Namibia	1950	1960	ILO, IHS, UN
Nepal	1900	1961	NSO, ILO, IHS, UN, Eg
Netherlands	1900	1900	OECD, GGDC, IHS, ILO, Smits *et al.* (1999)
New Zealand	1900	1900	OECD, ILO, IHS
Nicaragua	1900	1940	OLA, ILO, IHS, Eg
Niger	1900	1960	ILO, IHS, UN, Eg, Ma
Nigeria	1900	1960	Adeyinka *et al.* (2013), GGDC, UN, Eg, Ma
Norway	1900	1900	OECD, IHS, NAPP
Oman	1950	1993	ILO, UN
Pakistan	1900	1951	ILO, IHS, UN, Eg, Mc
Panama	1900	1940	OLA, ILO, Eg, IHS

Country	First obs.	First AES obs.	Sources
Papua New Guinea	1900	2000	ILO, UN, Eg, IHS
Paraguay	1900	1950	OLA, ILO, IHS, Eg
Peru	1900	1940	GGDC, ILO, OLA, IPUMS, IHS, Eg
Philippines	1900	1939	GGDC, ILO, IHS, Eg
Poland	1900	1900	OECD, GGDC, ILO, IHS, Deldycke *et al.* (1968)
Portugal	1900	1900	OECD, GGDC, IHS
Qatar	1950	1997	ILO, UN
Romania	1913	1913	ILO, IHS, UN, Good
Russian Federation	1900	1970	ILO, EF, Le
Rwanda	1900	1978	ILO, IHS, UN, Eg, Ma
Saudi Arabia	1950	1992	ILO, UN
Senegal	1900	1971	GGDC, UN, Eg, Ma
Serbia	1900	1900	ILO, IHS, My, Good
Sierra Leone	1900	1963	ILO, IHS, UN, Eg, Ma
Singapore	1920	1947	GGDC, IHS, Eg
Slovakia	1900	1900	OECD, GGDC, ILO, IHS, Good
Slovenia	1900	1900	OECD, GGDC, ILO, My, Good
Solomon Islands	1950	2009	NSO, UN
Somalia	n.a.	n.a.	
South Africa	1911	1911	GGDC, IHS
South Sudan	1950	2008	IPUMS, UN
Spain	1900	1900	OECD, GGDC; ILO, IHS
Sri Lanka	1900	1946	ILO, IHS, Eg
Sudan	1900	1956	IHS, UN, Eg, Ma
Suriname	1950	1973	ILO, UN
Swaziland	n.a.	n.a.	
Sweden	1900	1900	Schön and Krantz (2012)
Switzerland	1900	1900	OECD, ILO, IHS
Syrian Arab Republic	1900	1960	ILO, IHS, UN, Eg, Ka
Taiwan, Province of China	1905	1905	GGDC, ILO, IHS
Tajikistan	1900	1970	ILO, EF, Le
Tanzania, United Republic of	1900	1960	GGDC, UN, Eg, Ma
Thailand	1900	1937	GGDC, ILO, IHS, Eg
Timor-Leste	1950	2001	NSO, ILO, UN
Togo	1900	1981	ILO, IHS, UN, Eg, Ma
Trinidad and Tobago	1946	1946	ILO, IHS
Tunisia	1900	1956	ILO, IHS, UN, Eg, Ma
Turkey	1900	1927	ILO, IHS, Eg, Ka
Turkmenistan	1900	1970	EF, UN, Le
Uganda	1900	1991	ILO, IPUMS, UN, Eg, Ma
Ukraine	1900	1970	ILO, EF, Le
United Arab Emirates	1950	1995	ILO, UN

Country	First obs.	First AES obs.	Sources
United Kingdom	1900	1900	OECD, KLEMS, ILO, IHS
United States	1900	1900	OECD, KLEMS, ILO, IHS
Uruguay	1900	1950	OLA, IPUMS, IHS, Eg
Uzbekistan	1900	1970	ILO, EF, Le
Venezuela, Bolivarian Republic of	1900	1925	GGDC, OLA, ILO, IHS, Eg
Viet Nam	1920	1992	NSO, ILO, UN, Eg, IHS
Yemen	1950	1991	ILO, UN
Zambia	1900	1969	GGDC, UN, Eg, Ma
Zimbabwe	1900	1999	NSO, ILO, UN, Eg, Ma

References

Adeyinka, A., S. Salau, and D. Vollrath (2013): "Structural change in the economy of Nigeria." *International Food Policy Research Institute, Working paper No. 24*.

Almanach de Gotha (1910): *Annuaire Généalogique, Diplomatique et Statistique*. Justus Perthes, Gotha.

Almås, I. (2012): "International Income Inequality: Measuring PPP bias by estimating Engel curves for food." *The American Economic Review*, 102(2):1093–1117.

Åslund, A. (2001): *The myth of output collapse after communism*, volume 12. Carnegie Endowment for International Peace Washington, DC.

Bairoch, P. (1991): *Cities and economic development: from the dawn of history to the present*. University of Chicago Press.

Barro, R. J. (1996): "Determinants of economic growth: a cross-country empirical study." Technical report, National Bureau of Economic Research.

——— (2012): "Convergence and modernization revisited." *National Bureau of Economic Research Working Paper No.*

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

Caselli, F. (2005): "Accounting for cross-country income differences." *Handbook of economic growth*, 1:679–741.

Chen, X. and W. D. Nordhaus (2011): "Using luminosity data as a proxy for economic statistics." *Proceedings of the National Academy of Sciences*, page 201017031.

——— (2014): "A sharper image? Estimates of the precision of nighttime lights as a proxy for economic statistics." *Journal of Economic Geography*, page lbu010.

Chung, Y.-I. (2006): *Korea under siege, 1876-1945: capital formation and economic transformation*. Oxford University Press.

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

Costa, D. L. (2001): "Estimating real income in the United States from 1888 to 1994: Correcting CPI bias using Engel curves." *Journal of political economy*, 109(6):1288–1310.

Deaton, A. and B. Aten (2014): "Trying to Understand the PPPs in ICP2011: Why are the Results so Different?" Technical report, National Bureau of Economic Research.

Deldycke, T., H. Gelders, and J.-M. Limbor (1968): "La population active et sa structure." *Statistiques Internationales Rétrospectives*, 1.

Durlauf, S. N., P. A. Johnson, and J. R. Temple (2005): "Growth econometrics." *Handbook of economic growth*, 1:555–677.

Easterly, W. and S. Fischer (1995): "The Soviet economic decline." *The World Bank Economic Review*, 9(3):341–371.

Eggimann, G. (1999): "La Population des villes des Tiers-Mondes, 1500-1950." *Centre d'histoire economique Internationale de l'Universite de Geneve, Libraire Drozt.*

EICHENGREEN, B. (2008): *The European economy since 1945: coordinated capitalism and beyond*. Princeton University Press.

FAO (2012): "World food and agriculture." *Statistical Yearbook 2012*.

FEENSTRA, R. C., R. INKLAAR, AND M. TIMMER (2013): "The next generation of the Penn World Table." *National Bureau of Economic Research Working Paper 19255*.

GALOR, O. (2005): "From stagnation to growth: unified growth theory." *Handbook of economic growth*, 1:171–293.

——— (2011): *Unified growth theory*. Princeton University Press.

GOLLIN, D. (2010): "Agricultural productivity and economic growth." *Handbook of agricultural economics*, 4:3825–3866.

GOLLIN, D. AND R. ROGERSON (2014): "Productivity, transport costs and subsistence agriculture." *Journal of Development Economics*, 107:38–48.

GOOD, D. F. (1994): "The economic lag of Central and Eastern Europe: income estimates for the Habsburg successor states, 1870–1910." *The Journal of Economic History*, 54(04):869–891.

GOOD, D. F. AND T. MA (1998): "New estimates of income levels in central and eastern Europe, 1870-1910." *Von der Theorie zur Wirtschaftspolitik-ein österreichischer Weg*, page 147.

HAMILTON, B. W. (2001): "Using Engel's Law to estimate CPI bias." *American Economic Review*, pages 619–630.

HENDERSON, J. V., A. STOREYGARD, AND D. N. WEIL (2012): "Measuring economic growth from outer space." *American economic review*, 102(2):994–1028.

Herrendorf, B., R. Rogerson, and A. Valentinyi (2014): "Growth and Structural Transformation." *Handbook of Economic Growth*, 2.

IMF (2006): "Jamaica: Selected Issues." *IMF Country Report*, (06/157).

Jerven, M. (2013): *Poor numbers: how we are misled by African development statistics and what to do about it.* Cornell University Press.

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

Karpat, K. H. (1985): *Ottoman population 1830-1914: Demographic and social characteristics.* The University of Wisconsin Press.

Lagakos, D. and M. E. Waugh (2013): "Selection, agriculture, and cross-country productivity differences." *The American Economic Review*, 103(2):948–980.

Lampe, J. R. (1975): "Varieties of unsuccessful industrialization: the Balkan states before 1914." *Journal of Economic History*, pages 56–85.

Lewis, R. A., R. H. Rowland, and R. S. Clem (1976): *Nationality and population change in Russia and the USSR: an evaluation of census data, 1897-1970.* Praeger New York.

Lucas, R. E. (2009): "Trade and the Diffusion of the Industrial Revolution." *American Economic Journal: Macroeconomics*, 1(1):1–25.

Maddison, A. (2003): *Development centre studies the world economy historical statistics: historical statistics.* OECD Publishing.

——— (2010): "Statistics on world population, GDP and per capita GDP, 1-2008 AD." *Historical Statistics.*

Manning, P. (2010): "African Population: Projections, 1851-1961." In: K. Ittmann, D. D. Cordell, and G. H. Maddox (Editors), *The demographics of empire: the colonial order and the creation of knowledge*. Ohio University Press.

McEvedy, C., R. Jones, *et al.* (1978): *Atlas of world population history.* Penguin Books Ltd, Harmondsworth, Middlesex, England.

McGee, T. (1964): "Population: a preliminary analysis." In: W. Gungwu (Editor), *Malaysia: A Survey*, pages 67–81.

Minnesota Population Center (2008): "North Atlantic Population Project: Complete Count Microdata. Version 2.0 [Machine-readable database]." Technical report.

Mitchell, B. (1993): "International historical statistics: the Americas, 1750-1988."

——— (1998a): *International historical statistics: Africa, Asia & Oceania 1750-1993*. London.

——— (1998b): *International historical statistics: the Americas, 1750-1993*. New York: Stockton.

Myers, P. F. and A. A. Campbell (1954): *The Population of Yugoslavia*. US Department of Commerce, Bureau of the Census.

Nickell, S. (1981): "Biases in dynamic models with fixed effects." *Econometrica: Journal of the Econometric Society*, pages 1417–1426.

Nordhaus, W. D. (1996): "Do real-output and real-wage measures capture reality? The history of lighting suggests not." In: T. F. Bresnaha and R. J. Gordon (Editors), *The economics of new goods*, pages 27–70. University of Chicago Press.

Nuxoll, D. A. (1994): "Differences in relative prices and international differences in growth rates." *The American Economic Review*, pages 1423–1436.

Schön, L. and O. Krantz (2012): "Swedish Historical National Accounts 1560-2010." *Lund Papers in Economic History*, 123.

Smits, J.-P., E. Horlings, and J. L. Van Zanden (1999): *Dutch GNP and its components, 1800-1913*. Groningen Growth and Developement Centre.

Sobek, M., S. Ruggles, A. Trent, K. Genadek, R. Goeken, and M. Schroeder (2013): "Integrated Public Use Microdata Series, International: Version 6.2 [Machine-readable database]." *University of Minnesota, Minneapolis*.

The World Bank (2008): *Global Purchasing Power Parities and Real Expenditures. 2005 International Comparison Program*. The World Bank.

——— (2014): *Global Purchasing Power Parities and Real Expenditures of World Economies. Summary of Results and Findings of the 2011 International Comparison Program*. The World Bank.

United Nations (2012a): *World Urbanization Prospects, the 2011 Revision*. United Nations, New York.

——— (2012b): *World Urbanization Prospects, the 2011 Revision. Methodology*. United Nations, New York.

Wingender, A. M. (2014): "Structural transformation in the 20th century: Additional data documentation." *Available online at https://sites.google.com/site/asgerwingender/*.

——— (forthcoming): "Skill complementarity and the dual economy." *The European Economic Review*.

Young, A. (2012): "The African Growth Miracle." *Journal of Political Economy*, 120(4):696–739.

Zipf, G. K. (1941): *National unity and disunity*. Bloomington.