the review of income and wealth

Review of Income and Wealth
Series 64, Number 1, March 2018
DOI: 10.1111/roiw.12264

INDUSTRY WAGES ACROSS COUNTRIES AND OVER TIME: A NEW DATABASE OF MICRO SURVEY DATA

by Van Ha Le*

Faculty of Finance and Banking, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Jakob De Haan

De Nederlandsche Bank, Amsterdam, The Netherlands

CESifo, Munich, Germany

Faculty of Economics and Business, University of Groningen, The Netherlands

and

Erik Dietzenbacher

Faculty of Economics and Business, University of Groningen, The Netherlands

This paper presents a newly collected database on industry wages. The underlying data are micro datasets collected through nationally representative household surveys which are mostly conducted by national statistical agencies. In comparison with wage measures based on macroeconomic data sources, we find that industry wages based on micro survey data are more reliable. Furthermore, data from nationally representative micro surveys are becoming increasingly available, thus allowing for better coverage, especially of low-income countries. The database provides a reliable source of data for research on inter-industry wage structures and gender wage differentials, across countries and over time. It may also serve as a basis for further research on the determinants and implications of inter-industry and gender wage variations.

JEL Codes: J30, J31, J33

Keywords: industry wages, micro survey data, government wages

1. Introduction

This paper presents a newly collected database of industry wages for a large number of countries between the late 1980s and 2014. The main sources of this database are micro surveys. To the best of our knowledge, this is the first attempt to combine a diverse, scattered pool of survey data—which have been available for a large number of countries over the last two decades—into a unified database. As more data on micro surveys are being made available online through national data archives as well as cross-country databases, this database will be updated and expanded regularly in the future. The most recent version of the database can be downloaded online at https://microsurveydata.blogspot.com/.

*Correspondence to: Le Van Ha, Faculty of Finance and Banking, Ton Duc Thang University, Ho Chi Minh City, Vietnam (levanha@tdt.edu.vn).

Industry wages in this paper are defined as the mean nominal wages and salaries of all employees working in an industry. We determine wages for each of the one-digit level industries classified by the International Standard Industrial Classification of All Economic Activities, Revision 3 and 4 (ISIC-3 and ISIC-4). Different from the International Standard Classification of Occupations (ISCO), which classifies workers according to the skill content of their jobs, the ISIC describes the type of goods or services produced by the industry in which a person works.

Although survey data are far from perfect and may contain systematic (next to random) errors, validation studies have found that mean wages from a survey are unbiased estimates of the true figures, i.e. the average wages in the population (Bound *et al*., 2001).[1] The advantage of micro data is not only more accuracy but also better coverage. Data at the macro level on wages and employment are not available for many developing countries. If macro data are available, one may question how reliable macro-based estimates of average wages are. While the wage bill from national accounts data has been found to be unreliable in some countries (Bloem *et al*., 1998), it is in particular the measurement of employment that has been proven to be notoriously difficult in any country. This is due to the presence of part time, part year and casual workers (Heller and Tait, 1984; Schiavo-Campo *et al*., 1997; Lamo *et al*., 2012). In some countries, ghost workers pose a substantial problem on measuring employment at the aggregate level (Clements *et al*., 2010). Therefore, little can be said about the reliability government wages imputed by dividing the total recorded wage bill by total recorded employment. They may be unbiased, but they may also be seriously biased upward or downward, depending on the magnitude of the errors in the data for the wage bill and employment. Micro survey data, therefore, present a solution to facilitate academic research involving wages at the industry level.

As such, this dataset enables us to better understand the variation of wages across industries (i.e. inter-industry wages), both between and within countries. It will also provide a more reliable source of data to support empirical research on the determinants of such variations as well as their implications for other aspects of the economy. Micro survey data play an increasingly important role in social science research as well as policy analyses. However, most studies focus on analyzing wage data from surveys at the individual level while the potential of using survey data to study economic issues related to wages at a more aggregate level, such as industry sectors or the government sector, is still largely neglected.

Virtually all countries have conducted surveys on income and work in one form or another. Some countries, such as the U.S., Sweden and Argentina,

[1]Survey data may suffer from the problem of under-representativeness, especially when the surveys are conducted by individuals and institutions with limited resources. However, the surveys employed in our research are the results of national level institutions which are mandated to collect as comprehensive as possible surveys to support the policy-making process. In many developing countries, the surveys are conducted under the technical assistance of international institutions such as the World Bank or the International Labor Organization. There is therefore, a reason to believe that the problem of representativeness is dealt with in an appropriate manner.

maintain very long series of surveys with comprehensive data on income and work starting in the 1970s or even earlier. Other countries, such as Vietnam, Albania and Uganda, have only recently made household surveys a source of information for policy analysis thanks to the financial and technical support from international organizations. The results of such surveys are often made available via the World Bank or the International Labor Organization. Many countries offer free access to the original micro datasets for academic research via their own national data archives or international micro databases, such as the World Bank Living Standard Measurement Study or the Luxembourg Income Study.

Most countries nowadays have an online national data archive, and an increasing number of them allow free and instant access to their data after some simple registration. This availability enables us to build a database on industry wages across countries and over time. Our dataset was collected from 10 different international databases and a number of national data archives. The final database is a panel which consists of 1,834 per country/per year observations from 143 countries. On average, there are almost 12.8 observations per country, but the number of observations varies across countries.

The paper will proceed as follows. Section 2 presents the methodology to compute industry wages. Section 3 documents the data sources and the statistical operations applied to the data. Section 4 presents some detailed information about the database. Section 5 employs 101 micro survey datasets from 20 OECD countries to compare industry wages obtained from macro data and micro data. Section 6 gives some suggestions on how the database could be used and Section 7 concludes.

2. Methodology

2.1. *ISIC classification*

The ISIC system was introduced in 1958 by the United Nations (UN). Since then, it has been revised several times to meet the changes in industrial structure as well as the introduction of new economic activities. The first revision took place in 1968 (ISIC 2), the second revision was in 1989 (ISIC 3), and the latest revision (ISIC 4) was in 2008. Most micro survey data available by 2015 still follow the ISIC 3 classification. Some countries have employed the ISIC 4 classification, but parallel information based on ISIC 3 is often provided. The data prior to the 1990s, which mostly follow the ISIC 2 classification, are difficult to obtain and are incompatible with the data classified according to ISIC 3. For this reason, we focus on the late 1980s–2014 period.

Appendix A presents the ISIC 3 industrial structure. It is worth noting that the ISIC 3 system includes four levels of industries, from one-digit to four-digit. The larger the number of digits is, the more refined the industries are. UN member countries often adapt the ISIC system into their own local version such as the Nomenclature of Economic Activities (NACE) in Europe, the Vietnam Standard Industrial Classification (VSIC) in Vietnam, or the Industrial Classification for National Economic Activities (ICNEA) in China. At the one- and two-digit levels, the name of the industries is consistent across countries. However, the names

of the three-digit and four-digit level industries are more country-specific in order to describe more precisely the products of the corresponding economic activities.

Survey data from different countries vary. Some countries only report information at the one-digit level, while others collect data up to the four-digit level. We chose to present industry wages at the one-digit level for two reasons. First, focusing on the first level will ensure that the resulting data are more reliable. There are 17 one-digit, 60 two-digit, 159 three-digit, and 298 four-digit ISIC 3 industries. For some countries, however, the sample contains only about 1,000 surveyed individuals for which complete information on industry of work and wages is available. Extending the number of industries would then imply that some industries only have a few sampled individuals. A sample that is too small will result in estimated average wages that are unreliable. Second, many data at our disposal only contain information up to the ISIC 3 one-digit level, so that focusing on higher ISIC levels will lead to a significant reduction in the number of countries in the database.

Appendix B presents the ISIC 4 classification and its correspondence with the ISIC 3 system. Instead of 17 one-digit industries, the ISIC 4 consists of 21 industries, of which 10 industries are consistent with the ISIC 3 classification. These are: mining (C); manufacturing (D); construction (F); hotel and restaurants (H); financial intermediation (J); public administration (L); education (M); health (N); household activities (P); and extra-territorial activities (Q). Two ISIC 3 one-digit industries—agriculture and forestry (A) and fishing (B)—are combined into one ISIC 4 industry (named agriculture, forestry and fishing). The five remaining ISIC 3 industries are combined and reclassified into 10 ISIC 4 industries to reflect the introduction and development of new economic activities such as research, computer science and telecommunication.

As future waves of surveys will increasingly rely on the ISIC 4 system, we will present one time series for industries which are the same in ISIC 3 and ISIC 4. For the newly introduced and reclassified ISIC 4 industries, the data will only be available from 2008 onwards.

2.2. *Wage concepts*

The most comprehensive wage concept is wages and salaries defined in the 2008 System of National Accounts (SNA 2008, see United Nations *et al*., 2009). Wages and salaries include basic wages and salaries, overtime payment, and regular and irregular bonuses. Wages and salaries also include social contributions, income taxes, etc., payable to the employee (even if they are actually withheld by the employer for administrative convenience or other reasons and paid directly to social insurance schemes, tax authorities, etc., on behalf of the employee). Wages and salaries may be paid in various ways, including goods or services provided to the employees as remuneration in kind instead of, or in addition to, remuneration in cash (SNA 2008).

In the ideal case, a dataset of industry wages should employ the same concept of wages and salaries. However, the concept of wages and salaries used in surveys varies from one country to the other. Some countries report gross salary/pay, while others collect data on wages and salaries after taxes and social

contributions. Even for the same country, the concept of wages used in different surveys can differ over time, because surveys are reviewed periodically and changes in the questions on income may occur in subsequent rounds.

Given this situation, we did not seek a common definition of wages in our database. Instead, we retain the concepts of wages used in each underlying survey. As a result, the industry wages reported in our database vary across countries and sometimes over time as well. A variable named WAGE_CONCEPT is included in the database to outline the concept of wages and salaries used in a particular survey.

Comparing industry wages across countries may not be appropriate when the wage concepts used in the underlying surveys differ from each other. Similarly, the trend analysis of industry wages series may be disrupted if the wage concept within a country changes from one year to the other. However, the differences in wage concepts used will be less of a problem in studies that focus on relative differences in wages and salaries across sectors. This is because the wage concepts have always been applied consistently to all sectors within one survey. If components of wages and income tax policies do not differ across industries, the relative position of industries will not depend on the use of a particular wage concept.

2.3. *Time basis*

Time basis refers to the reference period of the wages and salaries (e.g. dollars per year or per hour). Similar to the wages and salaries concept used, also the time basis varies between and within countries. Within the pool of original micro datasets that we have collected, eight different time bases have been used, namely yearly, 6 monthly, 3 monthly, monthly, fortnightly, weekly, daily and hourly. The time basis may even vary within one survey. If that is the case, we convert the reported income into the most commonly used time basis in the respective survey and report the wages according to that basis. The variable TIME_CONCEPT in the database provides the time basis for the reported wages and salaries.

Most surveys provide information on weekly hours worked. To create a measure consistent in terms of time coverage, we also compute annual wages by multiplying the reported wages by 52 or 12 if they are weekly or monthly, respectively. In case a country reports only hourly wages, the annual figure is computed using the data on hours worked. For example, if an individual works 38 hours per week and receives an hourly payment P, his/her annual wages are computed as the hourly payment times the number of hours worked per week times 52, i.e. $P\times$ *number of hours worked*$\times 52$. The hourly wages take differences in working hours across individuals into account. However, survey data on hourly wages are less accurate than data on weekly, monthly and yearly wages (Bound *et al*., 2001). For that reason, we choose to present the annual wages across countries, next to the original measure in the survey.

2.4. *Wages by gender*

Besides producing reliable estimates, micro survey data also allows the distinction of wages by gender. Gender is readily available in almost all socio-

economic surveys. As a result, we present industry mean wages for female and male workers separately, next to the mean wages for all employees.

2.5. *Estimation procedures*

The estimation procedures consisted of two steps. First, surveyed individuals with a positive amount of income from the main job are classified into one of the 17 ISIC 3 (see Appendix A) or 21 ISIC 4 industries (see Appendix B). Second, the wages of each industry are estimated as the mean reported wages from the main job of all individuals within that industry. More specifically, the wages of industry i in a given country at time t were estimated as:

$$W_{it} = \frac{1}{\sum_{j=1}^{n_{it}} p_{itj}} \sum\nolimits_{j=1}^{n_{it}} \left(p_{itj} \times w_{itj}\right)$$

where W_{it} are the wages of industry i at time t, w_{itj} are the reported wages of the surveyed individual j in industry i at time t, n_{it} is the number of surveyed individuals in industry i at time t, and p_{itj} is the sampling weight of individual j in industry i at time t. These sampling weights are provided in each dataset to account for the differences in sampling probability between individuals. When the sampling process is completely random, individuals within each country will have an equal probability of being surveyed. In that case the weight $p_{itj}=1$ and the expression reduces to $W_{it}=\sum_{j=1}^{n_{it}} w_{itj}/n_{it}$. However, in practice, most surveys are stratified along several dimensions and the probability of being interviewed is different from one individual to another. The sampling weight p_{itj} is included to account for this difference.

3. Data Sources

3.1. *Original micro survey datasets*

The main sources of the original survey datasets that we collected include 10 international databases together with a number of national data archives. Major international databases used include the World Bank Living Standard Measurement Study (LSMS), the European Community Household Panel (ECHP), the European Union Statistics on Income and Living Conditions (EU-SILC), and the Luxembourg Income Study (LIS).[2] National data archives are online databases that offer original micro survey datasets conducted in each country and over time, such as the Current Population Survey (CPS) from the United States. Access to these data is often free for academic research purposes (under certain conditions). In total, we were able to collect around 1300 original datasets between the late 1980s and 2014, which was our starting-point.

Our database relies mainly on three types of surveys namely household surveys, labor force surveys and establishment surveys. While the units of survey in

[2]The Structure of Earnings Survey (SES) of Eurostat is conducted only every four years. In years for which SES data are available, the EU-SILC data are also available. As EU-SILC data are available for more years we use this source.

the first two types of surveys are randomly chosen individuals, the units of survey in the third type are randomly chosen firms, i.e. the establishments. There is a large number of studies that confirm the reliability of mean self-reported wages (Bound *et al*., 2001). However, there are not many studies comparing the reliability of wages collected in establishment surveys. In a recent publication, the African Development Bank reports that there is some difference between the two types of surveys but the difference is not significant (AFDB, 2012).

For some countries, the information on the industry (according to the ISIC classification) in which respondents work is missing. Also, when the industry of work is available, some surveys do not report data on wages or fail to distinguish income from work and income from other sources. The accompanying manuals and questionnaires were carefully examined to ensure that the necessary information is correctly recorded and only surveys with data of good quality were retained. That is, datasets without clearly defined variables on income from work and industry of work were not processed. Datasets with only subnational coverage or containing too few working individuals were also omitted because the resulting sample was unreliable.

Most datasets report data on a monthly basis. However, data on bonuses and overtime payment as well as earnings in kind are—when available separately—often reported on a less regular basis, such as 3, 6 or 12 months. In that case, we converted these incomes to a monthly basis and added the resulting income to the monthly wages to obtain the monthly earnings data. The obtained earnings data can either be in gross terms or in net terms. The resulting earnings figures were considered as the final reported wages and salaries of the surveyed individuals and were used to estimate the industry wages and salaries.

In total, we processed 958 datasets from 80 countries, resulting in 760 observations. The number of datasets processed is larger than the number of per country/year observations because some countries provide data on a quarterly or biannually basis. For these countries, the final data in a given year is the average of the data computed from all datasets available for that year. For example, the annual industry wages for the UK in 2015 are the average of the mean industry wages estimated from all the four quarterly datasets obtained from the UK Data Service database.

3.2. *Secondary data on industry wages*

Secondary data come from international databases that provide country data on survey-based estimated wages for ISIC industries. These databases often aim at a comprehensive coverage of different aspects of labor and income. Industry wages are thus often provided together with data on the population structure, the income distribution, and the labor share. The International Labor Organization is the leading international institution in this respect, but also the World Bank has, in cooperation with local institutions, created a database on income and distribution in Latin America and the Caribbean. One drawback of these databases is their bias towards developed countries. For many countries, especially low-income countries, wage data are not included in the ILO database even though their household survey data are available. Another problem of using secondary sources for data on industry wages is that some industries, especially the public

administration industry, are often not included. Wages by gender are also not reported in many cases.

In constructing our dataset on industry wages, we have used information from three existing databases, namely SEDLAC, the International Labor Organization's LABORSTA and its successor, ILOSTAT.

SEDLAC is a database that provides several socio-economic statistics for 25 Latin American and Caribbean countries. It is the result of the cooperation between the Center for Distributional, Labor and Social Studies (CEDLAS) of the University of La Plata and the World Bank's Latin America and the Caribbean Poverty and Gender Group (LCSPP). SEDLAC provides data on mean net monthly earnings and hourly wages based on household budget survey data. Instead of reporting data by ISIC 3 industries, SEDLAC provide average wages for 10 sectors. These sectors are primary activities; low-tech industries; high-tech industries; construction; commerce; utilities and transportation; skilled services; public administration; education and health; and domestic servants.

Appendix C shows details on the relationship between the 10 SEDLAC sectors and the 17 ISIC 3 industries. The wages for construction (industry F in Appendix A) can be obtained directly from the SEDLAC database. The wages for manufacturing (industry D) can be obtained as the weighted average of the wages of the low-tech and hi-tech sectors (the detailed formula is presented in Appendix C). Each of the other 7 SEDLAC sectors combines some ISIC 3 one-digit industries together. We report the wages for these sectors as they are reported in the SEDLAC database. 292 observations on wages and salaries have been collected from this database.

LABORSTA provides data on average wages for ISIC 3 industries for a large number of countries. The ILO Department of Statistics sends questionnaires to the corresponding government agency in each country/territory to collect data on various labor statistics. With respect to industry wages, the corresponding agency is requested to provide data on gross earnings by ISIC industries. When data on average earnings is not available, data on the average wage rates is provided instead. A detailed list of components of remuneration to be included in each wage concept is also included in the questionnaires. When the questionnaires are returned to the ILO, the data are checked for their consistency and made freely available online at http://laborsta.ilo.org/(Hua, 2008). More recently, the LABORSTA database has been replaced by ILOSTAT, which provides additional data on earnings from 2009 to 2014. Data from ILOSTAT are classified into both ISIC 3 and 4 industries, depending on country/year.

We included LABORSTA and ILOSTAT into our database. For the period from the late 1980s to 2014, we were able to collect 1,313 observations from 115 countries and territories. However, the data are not evenly distributed between countries. For some countries there is a very long data series while for many other countries there is only one observation.

4. Results

The final database includes three separate datasets corresponding to the three data sources that have been used, namely the SEDLAC, the ILO databases

TABLE 1
Distribution of the Database by Regions and by Source of Data

Region	Final dataset	Original micro-data	Secondary Data From SEDLAC and ILO
Latin America & Caribbean	415	221	376
Central Asia	100	4	98
Middle East & North Africa	112	29	99
East Asia & Pacific	232	36	201
Europe	787	378	666
North America	59	35	33
South Asia	14	7	7
Sub-Saharan Africa	115	50	52
Total	1834	760	1532

and the original micro datasets that we have collected. The online Appendix D provides summary information on the data source for each country in all years covered in the database. The website http://microsurveydata.blogspot.com/ provides the data sources used as well as the processing steps applied to get each observation in detail.

In addition, there is a unified dataset that combines these three datasets. Sometimes we have more than one observation for a particular country-year. In that case we retained the entry that provides more comprehensive data, both in terms of the number of industries covered and the wage concepts used. When two different sources provide the same wage concept and time basis, the longer series has been retained. For France, for example, the ECHP provides data for the average gross monthly earnings in cash for 1994-2001 and the EU-SILC for gross yearly earnings in cash and in kind for 2003–10. At the same time, the ILO provides data on gross monthly earnings in cash for 1999–2002. So there are multiple entries for France in 1999, 2000 and 2001. We chose to include data for France in the 1994–2001 period from the ECHP because this ensures continuity of the data series.[3]

As shown in Table 1, the final combined dataset consists of 1,834 observations for 143 countries and territories. For Europe we have the largest number of observations, 787. There are 415 observations from Latin America and Caribbean; 232 observations from East Asia and Pacific; 115 observations from Sub-Saharan countries; 112 observations from the Middle East and North America; 100 observations from Central Asia; 59 observations from North America and 14 observations from South Asia. Figure 1 presents the number of observations per year, distinguishing between OECD and non-OECD countries.[4] There are less than 20 countries included in the database in the early 1990s.

[3]As suggested by one of the referees, we have constructed several indicators of public administration wages relative to other industries such as manufacturing and education (i.e. dividing the public administration wages by the corresponding wages in the respective industries) for countries for which we have several data sources. The correlation coefficients are relatively high (ranging from 0.61 to 0.7).

[4]A country is classified as an OECD member from the year it is admitted to the OECD. The increase in the number of OECD countries may therefore reflect that data from more OECD countries have become available and/or that OECD membership has increased.

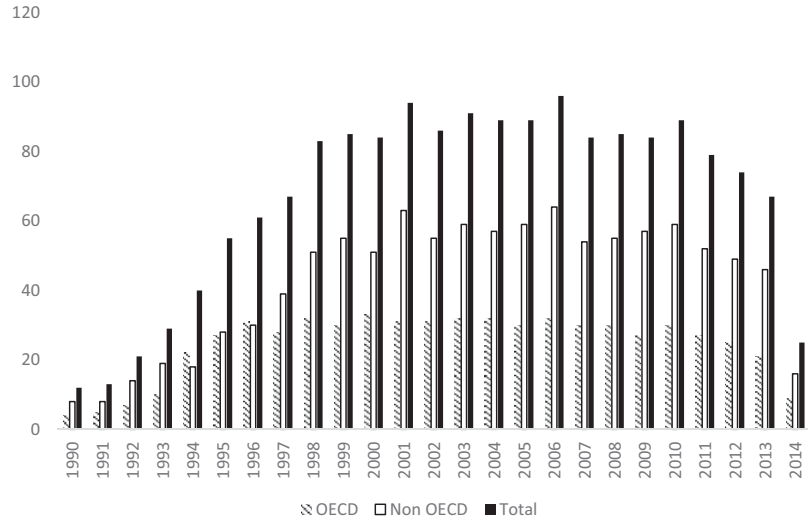

Figure 1. The number of countries included in the database by year and by OECD status

However, the number of countries increases gradually and reaches between 60–100 countries in the 1996–2013 period. In 2014, the number of countries appears to be smaller because of the lag in data dissemination.

There are 17 ISIC 3 and 21 ISIC 4 industries, and for each industry we report average wages using the original time basis and average wages on an annual basis. There is also information on annual wages for females and males in each industry.[5] Finally, information is provided on the sample size for each observation estimated based on the original micro datasets that we have obtained.[6]

Figure 2 shows the number of observations per country on the horizontal axis and the number of countries on the vertical axis. The figure shows that 15 countries have only one observation, and 43 countries have 20 observations or more. On average, there are 12.83 observations per country.

Table 2 lists the types of surveys underlying our database. More than 43 percent of the data are from labor related establishment surveys[7], almost 39 percent are from household budget surveys, and about 11 percent of the data are from labor force surveys. Less than 6 percent of the data are from other sources, such as administrative reports, tax records and population censuses. The variable SOURCE in the database gives details about the type of survey for each observation in the database.

[5]To support further research on issues relevant to industry wages, the database also provides measures of labor share by sex and by industry and median wages by sex and by industry whenever the data is available.

[6]To facilitate further research, we will also add measures of income inequality; and mean and median wages, labor share of the government sectors in the next version of the database.

[7]Labor related establishment surveys are conducted at the establishment level, while labor force surveys collect information by interviewing individuals. In some countries, the labor related establishment surveys exclude small firms having a number of employees below a certain threshold. The most commonly applied threshold is five but sometimes it is 10 or 20.

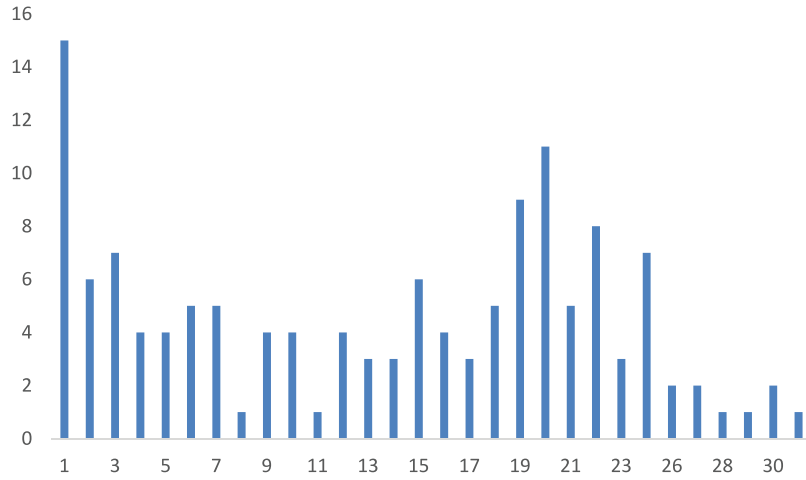

Figure 2. Number of countries by number of observations [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2

Types of Surveys

Sources of data	Frequency	Percent
Labor related establishment survey	789	43%
Household budget survey	711	39%
Labor force survey	193	11%
Administrative reports	48	3%
Population census	38	2%
Insurance records	34	2%
Tax records	9	0%
Total	**1834**	**100**

There is also a WORKER_COVERAGE variable to record information on the worker coverage in our database. Four different levels of worker coverage are used in the data, namely: total coverage; employees; salaried employees; and wage earners. Salaried employees receive a payment that is independent of the number of hours worked and wage earners are low-skilled workers whose wages depend on their hours worked. Employees covers total paid employment, both by salaried employees and by wage earners while total coverage includes employees and self-employed people. The worker coverage in our database is consistent, which refers to employees in most of the cases.

Table 3 presents the composition of our database with respect to the concept of wages that is used. About 81 percent of the cases report data on a monthly basis, out of which 53 percent refer to gross monthly earnings and 28 percent to net monthly earnings. Earnings per year, earnings per week and earnings per hour are the next three major categories, which account for 8 percent, 4 percent and 3 percent, respectively. Less than 3 percent of the data are based on other concepts

TABLE 3
WAGE CONCEPTS

Wage Concepts	Frequency	Percentage
Earnings per month	965	53%
Net monthly earnings	517	28%
Earnings per year	154	8%
Earnings per week	74	4%
Earnings per hour	56	3%
Earnings per day	31	2%
Net earnings per year	13	1%
Wage rates per month	4	0%
Total	**1834**	**100**

of wages, such as monthly salary, earnings per day or wage rate per day. The variable WAGE_CONCEPT in the database gives details about this issue.

The database also provides information on sex and therefore allows investigating wage differentials between male and female employees. Figure 3 presents the gender wage-gap for some industries and for the economy as a whole between the late 1980s and 2014. Each box graph shows the distribution of the ratio between the mean wages for men and the mean wages for women. The closer the value of this ratio is to 1, the smaller the wage gap is. Consistent with other studies on wage inequality (Hotchkiss and Pitts, 2007; Beaudry and Lewis, 2014;), Figure 3 shows that women earn less than men across all industries.

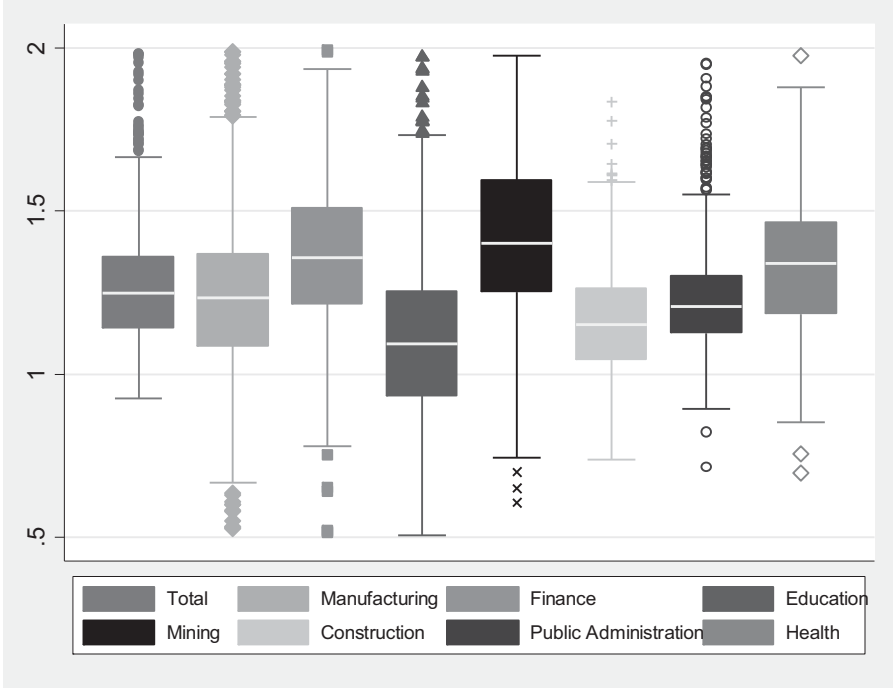

Figure 3. Gender wage gap at the industry level

5. The Possible Bias of the Macro Data Approach

As pointed out in the Introduction, using macro data may be an alternative for using micro data. In this section we compare micro and macro data, focusing on government wages, to explore the reliability of macro data in more detail. The macro data approach computes the average government wages as the ratio of the government wage bill to total government employment. Potential biases therefore occur both in the numerator and in the denominator. First, the government wage bill may not be recorded correctly (or inconsistently across time and space). Many non-monetary benefits—such as travel, housing and other allowances—can be recorded as spending under other goods and services. In a similar fashion, payments to employees in government projects can be recorded as capital spending and outlays for temporary workers are in some countries classified as use of goods and services (Clements *et al*., 2010). In general, a number of spending categories should be included in the government wage bill but they are often mistakenly excluded (i.e. recorded somewhere else).

Second, creating government employment data that are consistent over time and across countries is notoriously difficult. Issues to cope with include full time versus part time workers and full year versus part year employment. Some employees work full time for the whole year but others may work part time and only for a fraction of the year. In addition, casual workers—who work on contingent government projects—may account for a substantial proportion of government employment during some periods (Klitgaard, 1989). The key problem is that there is no agreed international standard on how to combine these different types of government employment into a single figure. An issue that applies to some developing countries is that government employment is inflated by fictitious workers who only exist in the books (Clements *et al*., 2010).

The imputed average annual government wages can thus be biased upward or downward, depending on the precision of the statistics in a country. The changes in annual government wages within a country may simply reflect changes in the statistical methods employed rather than actual changes. When such measurement errors are not random, empirical conclusions can be rendered unreliable.

However, also survey data in general and survey data on wages and salaries in particular suffer from measurement errors. A large body of literature on survey data validation has been devoted to analyzing the characteristics of errors in the reported wages (see Akee, 2011 and references cited therein). These studies compare survey data on gross earnings with administrative data such as social security records or employer records. Both for developed and developing countries, the survey errors have been found to center around zero and to be mean reverting. Mean reversal occurs when people at the higher end of the income distribution underreport their income while people at the lower end of the distribution over-report their income. Based on this finding, let the self-reported survey wages and salaries of individual j be given by X_j. We assume

$$X_j = X_j^* - \delta\left(X_j^* - \bar{X}^*\right) + \varepsilon_j$$

where X_j^* gives the true wages and salaries for surveyed individual j, $\bar{X}^*$ is the population mean of true wages and salaries, and ε_j is white noise centered around 0. The term $\delta\left(X_j^* - \bar{X}^*\right)$, with $\delta > 0$, captures the mean reversal of the errors. Suppose that a nationally representative sample of n government employees is collected via a survey. The sample mean of the government wage rate is $\bar{X} = \sum_{i=1}^{n} X_i/n$ and its expected value is

$$\begin{aligned} E(\bar{X}) &= \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[X_j^* - \delta\left(X_j^* - \bar{X}^*\right) + \varepsilon_j \right] \right\} \\ &= \frac{1-\delta}{n} \sum_{i=1}^{n} E\left(X_j^*\right) + \delta\bar{X}^* + \frac{1}{n} \sum_{i=1}^{n} E(\varepsilon_j) \\ &= \frac{1-\delta}{n} n\bar{X}^* + \delta\bar{X}^* = \bar{X}^* \end{aligned}$$

where we have used that the sample is representative, i.e. $E\left(X_j^*\right) = \bar{X}^*$, and that the errors center around zero, i.e. $E(\varepsilon_j) = 0$. The sample mean of the government wages in the survey is therefore an unbiased estimator of the population mean. This theoretical result, which rests on certain assumptions, has been corroborated empirically in various survey validation studies (see Bound *et al.*, 2001, for a survey). Most studies report that the mean wages and salaries differ just a few percent between the survey and the administrative records (e.g. Akee, 2011). The finding that survey validation studies always report that the mean wages and salaries in survey and administrative data are close to each other suggests that mean wages from survey data are an unbiased estimator of the mean wages from administrative data. Therefore, we use the average government wages from household surveys as a benchmark to validate the government wages computed from macro data sources.

In most countries, survey data often provide only some components of wages and salaries, such as net earnings, basic wage rates or a combination of basic wage rates and some other components like overtime payment or bonuses. In this section we use the European Union's "Statistics on Income and Living Conditions" (EU-SILC) database and the Current Population Survey (CPS, March Supplementary files) from the United States. The EU-SILC database results from a harmonized framework by the 27 EU member countries plus Croatia, Iceland, Norway, Switzerland and Turkey to collect comprehensive data on income and living conditions (Clemenceau and Museux, 2007). We have used the cross sectional EU-SILC data from the first seven rounds of the project, which took place between 2004 and 2010. In total, 177 datasets from 29 countries were obtained. Data for Croatia, Switzerland and Turkey are not available because these countries only participated in the most recent rounds. The survey in a given year t, provides for each individual the following information: the industry of work in year t according to the International Standard of Industrial Classification of Economic

Activities, Revision 3 (ISIC 3); the total wages and salaries in cash and in kind from work in year t-1; and whether an individual had a second job in year t-1. Using information that is available on a monthly basis, it is also known whether an individual changed jobs during the last 12 months. The wage concept used in the EU-SILC surveys is comparable with the definition of wages and salaries as used in macro databases.

Our calculation of the wages in public administrations involved several steps. First, only individuals who work for the public administration industry are included. Second, public administration employees who report that they have changed their job since last year were dropped because their reported wages and salaries were earned in their previous job. Third, individuals who report that they had a second job were also dropped because their wages and salaries may contain income that is not earned in public administration. Fourth, some countries do not report data on whether an individual had a second job. If that information is not available, the whole dataset of that year is dropped. As a result, only 134 datasets in the EU-SILC database have been used. In the final step, our micro-based estimates of annual government wages are obtained as the weighted averages of the gross annual earnings of the remaining individuals.

The CPS is conducted monthly by the U.S. Department of Labor and is representative for the whole population in the United States. Different from the EU-SILC data, the CPS only provides data on gross weekly earnings from the main job. For our study, we retain only individuals who work for the public administration and estimate the mean weekly wages for this subsample. The annual government wages are computed as the weekly wages multiplied by 52. This practice results in an underestimation of the annual wages if public administration employees enjoy non-regular earnings that are not captured in the weekly earnings.[8]

One advantage of both the EU-SILC and the CPS is their representativeness. The EU-SILC gives substantial freedom to national statistical offices to design their own sample. For each country, four different nationally representative subsamples are surveyed in each year. In the following year, one subsample is dropped and replaced by a completely new subsample. Each subsample is therefore included in the survey for four years. In a similar manner, the CPS sample also includes four rotating subsamples and one subsample is replaced in every month. This design implies that the overall sample is representative, updated, and the sampling errors are likely to be close to zero because each country/year dataset includes 4 independent nationally representative samples.

Our macro data are taken from the OECD Structural Analysis (STAN) database. The STAN database provides data on the wage bill and employment for ISIC 3 industries in 33 OECD countries from the early 1990s onwards and is updated regularly. The data come from the national accounts of the member countries. The macro average annual government wages (or STAN macro wages) are computed by dividing the wage bill by the number of employees in public administration. All countries included in the STAN database use head counts as the measure of employment. The macro data are thus in line with the micro data

[8]As will be made clear later, this practice does not undermine our conclusion for the case of the U.S.

TABLE 4
SUMMARY OF THE RATIO OF MACRO TO MICRO

Country	Observations	Mean	Standard Deviation	Min	Max
Austria	7	0.99	0.05	0.93	1.09
Belgium	5	1.00	0.04	0.96	1.04
Czech Republic	6	1.20	0.03	1.16	1.23
Denmark	7	1.30	0.07	1.20	1.40
Estonia	7	1.19	0.10	1.09	1.36
Finland	1	0.82	–	–	–
France	5	1.09	0.05	1.02	1.15
Germany	5	0.94	0.02	0.91	0.97
Greece	1	1.03	–	–	–
Hungary	5	1.44	0.08	1.36	1.54
Ireland	3	1.24	0.02	1.22	1.25
Italy	4	1.06	0.03	1.04	1.10
Luxembourg	7	1.10	0.02	1.08	1.14
Netherland	6	0.89	0.02	0.87	0.92
Poland	4	1.07	0.02	1.05	1.11
Portugal	1	1.13	–	–	–
Slovakia	6	1.81	0.20	1.64	2.09
Slovenia	6	1.09	0.01	1.07	1.11
Spain	3	0.96	0.04	0.94	1.01
United States	12	0.93	0.03	0.90	1.01

in the sense that no distinction is made between full time and part time workers. Some countries, however, also provide full time equivalent units. For these countries, we are therefore able to compute the corresponding full time equivalent annual government wages.

STAN data are in the national currency while all EU-SILC data are in euros. To make a comparison possible, the STAN data on the wage bill are converted to Euros using the average annual exchange rates as obtained from Eurostat. This is done for all countries not in the Euro area, except the United States (for which both macro and micro data are in current U.S. dollars).

The final sample consists of 101 observations from 20 countries. The reduction of the sample size is due to the fact that some countries are included in the STAN database but not in the EU-SILC database, and vice versa. All observations fall in the period 2003–2009, except the annual series for the United States which spans 1998–2009. For most countries we have more than one observation except for Finland, Greece and Portugal.

To ease the comparison of the macro and micro wages and highlight the bias of the macro data, we divide the macro government wages by the micro government wages. The resulting ratios are Table 4.

Although the average micro government wages are unbiased estimates of the true population mean of the government wages, they still contain sampling errors. When the sample—for a given country in a given year—would be repeated a large number of times, the average micro government wages will follow a normal distribution with the true average government wages as its mean. If the macro approach uses the same definition for the government wages and if the macro wages do not include measurement errors, they also reflect the true population

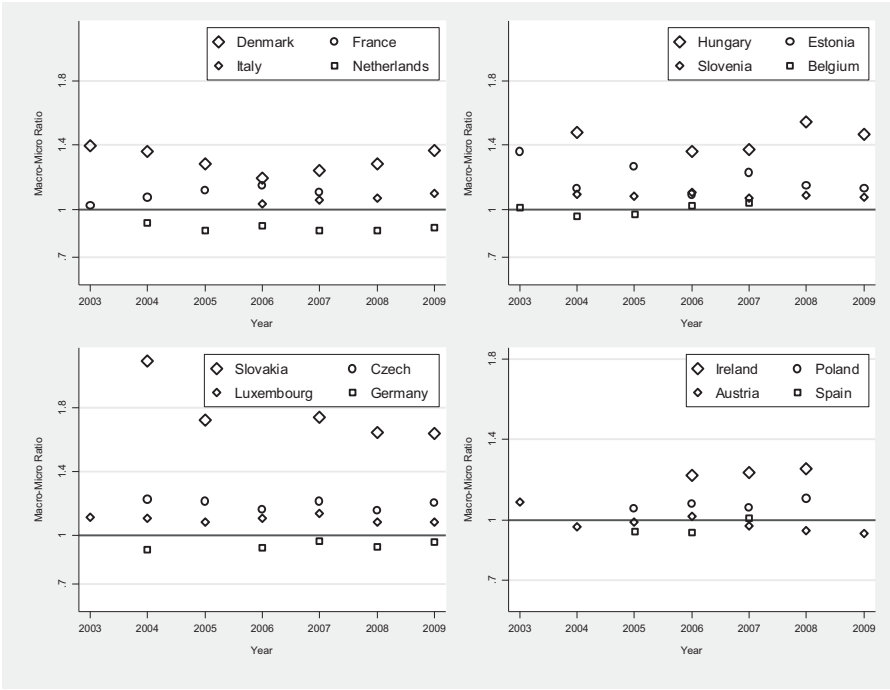

Figure 4. The distribution over time of the ratio of the macro to the micro public administration wages

mean. In that case, the macro and micro wages must—for individual years—be approximately the same and the macro-micro ratios will oscillate around 1. If the macro-micro ratio is consistently larger (smaller) than 1, the macro data approach overestimates (underestimates) the true population mean of the government wages. This is because the micro data approach yields an unbiased estimate (i.e. only contains random sampling errors). If macro and micro data use the same wage definition, a macro-micro ratio that is consistently larger (or smaller) than 1 suggests the occurrence of systematic measurement errors in the macro approach. For example, it may be the case that a country's statistical practices lead to errors in the measurement of the wage bill and/or total government employment. Because statistical practices are often in place for some time, the error is likely to remain from one year to another.

The results in Table 4 show that the macro data approach performs relatively poor for most of the 20 listed countries. It only produces good estimates of government wages for Austria, Belgium, and—to a lesser extent—Spain and the U.S. For these countries we find that the smallest observation is below 1 and the largest above 1. For the remaining countries we find that the macro wages are persistently too high for 11 countries and persistently too low for two countries (Germany and the Netherlands), while for three countries (Finland, Greece and Portugal) we only have one observation.

For the U.S., we find a macro-micro ratio that is smaller than 1 in 11 out of the 12 years of observation (the exception is 1.01 in 2007). It should be emphasized that the micro data for the U.S. are not annual figures. Instead, the annual government wages have been computed from weekly gross earnings that were multiplied by 52. This practice may fail to include non-regular payments, in which case the micro government wages have a negative bias. If the macro approach contains no measurement errors, the imputation procedure would provide the correct government wages. In that case, we would expect to find that the calculated micro wages are smaller than the calculated macro wages. Clearly, this is not the case. The macro government wages are lower than the micro wages, which indicates that the macro data approach underestimates the government wages for the U.S.

Figure 4 graphs the macro-micro ratio over time and illustrates the persistence of the bias of the macro data approach. For instance, the ratio for Hungary fluctuates around 1.5 suggesting that the macro data approach overestimates the government wages by about 50 percent. In a similar manner, the ratio for the Netherlands oscillates around 0.9 suggesting that the macro wages underestimate the government wages by about 10 percent. The worst case is Slovakia, which has all data points above 1.6. We have also extended our comparison to other ISIC 3 industries and find that the inadequacy of the macro data approach for deriving the average wages is also present in industries other than public administration (results available on request).

6. Suggestions for Future Usage of the Database

Most studies on cross-country industry wage differentials thus far use macro data, mainly coming from the national accounts (Gittleman and Wolff, 1993; Gaetano, 2011; Genre *et al*., 2011; Hartwig 2011). Our database offers an alternative and more reliable basis to revisit empirical findings which are based on macro data. More importantly, this database will open new opportunities for research on different relevant economic issues in a large number of countries.

First, this database enables research—with reliable data—to further our understanding of the structure of industry wages, both within and between countries. Although the concept of wages varies between countries and sometimes within one country, the wage concept is consistently used within one survey. As such, our database will be suitable for analyzing relative wages of different industries within a country. To the extent that taxes and components of wages other than the basic wages and salaries rate are proportional to total earnings, research concerning the inter-industry wage structure will not be likely to suffer from any loss of precision.

For illustrative purposes, Figure 5 shows the ratio of the average wages of some industries to the average wages of the whole economy. Each industry is indicated by its ISIC 3 one-digit level. The figure reveals that the inter-industry wage structure is relatively stable over time. Previous studies on inter-industry wages focused on industrial countries due to lack of data for developing countries. Our database enables the study of the inter-industry wage structure on a worldwide basis and to examine differences between different sets of countries.

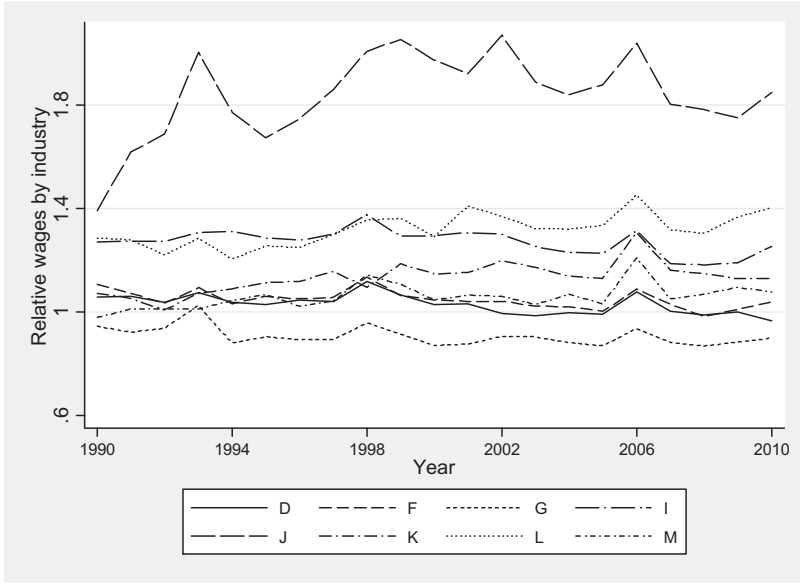

Figure 5. The inter-industry wage structure

Second, our database can be used to study the determinants of the inter-industry wage structure for a large number of countries. Based on a sample of eight Euro area countries over the 1991–2002 period, Genre *et al*. (2011) find that human capital, firm characteristics as well as socio-cultural and institutional factors play a major role in explaining inter-industry wage differentials. Our database will help extending this line of research to other countries, especially developing countries for which data availability is often a severe problem. This database will also provide a basis for empirical studies on the impact of globalization, especially the possible impact of international trade and capital flows, on the relative wage position of tradable industries, such as manufacturing and mining, and non-tradable sectors, such as education, public administration and health.

Third, our database offers a new opportunity to study the impact of the inter-industry wage structure on different aspects of the economy. There is a large body of literature on the relationship between wages and workers' performance. According to the efficiency wage theory, high-paid workers will be less likely to shirk and more willing to exert effort to fulfill their duty. This will improve the profitability of their firms and—at the aggregate level—the efficiency of the whole industry. This line of argument has been applied to investigate the relationship between government wages and corruption (Van Rijckeghem and Weder, 2001) and the quality of government (La Porta *et al*., 1999). Similarly, our database can be employed to extend the analysis to the relationship between wages and service quality in other industries, such as education, and health and social work.

Fourth, our database can be used to infer government wages. The database gives the average wages of the public administration (Industry L in Appendix A, and Industry O in Appendix B). This narrow definition of the government

excludes government activities in areas such as education and health, while it includes some private sector activities. However, based on a sample of 194 surveys from 43 countries (for which we have data to classify government and non-government employment in the public administration industry) it can be shown that the wages of the public administration industry are a good proxy for government wages. The private component in public administration is very small and estimated government public administration wages are very close to total public administration wages (detailed results are available on request).

Finally, our database can be employed to study issues relating to labor's share of income or the Baumol's disease and the wage leadership of the government sector. The database will also lay the foundation for studying aggregation biases in aggregate wages and the cyclical responsiveness of wages in general.

7. Conclusions

Lack of data has been a stumbling block in empirical research on industry wages at the cross-country level. Most studies so far relied on data from national accounts and were mostly confined to a group of developed countries with relatively good data at the macro level. This paper introduces a micro-based database which covers a large number of developing—next to developed—countries. It thus makes it possible to extend analyses to countries which have so far been neglected due to lack of data. As more survey data become available, the database will be updated regularly in the future. The next version of the database will also provide data on additional variables such as income distribution (1^{st}, 2^{nd} vs. 8^{th} and 9^{th} deciles), and education (where possible) to help further research on relevant issues.

The database consists of 4 datasets. The first three datasets correspond to the data collected from the ILO, the SEDLAC, and the original micro survey datasets we have collected from nine international databases and several national archives. The fourth dataset combines these three datasets into a unified dataset. When a particular country-year has more than one observation, this unified dataset includes the observation which is based on the more comprehensive concept of wages. When two observations are based on the same wage concept, the observation from the series which is available for the longest time span is included.

Although the concept of wages, time basis and worker coverage vary from one country to another, our database is suitable for studying relative wages between industries within the same country and between countries. This is because the concepts of wages, time basis and worker coverage used in the database are—within each country—consistent over time. Therefore, analyses based on relative wages are not likely to suffer from any loss of precision. Authors who are interested in comparing the level of industry wages between countries or analyzing the trend of wages over time should consult the online explanatory notes as well as the concept variables which are provided in the database to ensure that data are comparable.

References

African Development Bank (AFDB), "Labour Force Data Analysis: Guidelines with African Specificities," Statistics Department, Tunis, 2012.

Akee, R., "Errors in Self-reported Earnings: The Role of Previous Earnings Volatility and Individual Characteristics," *Journal of Development Economics*, 96, 409–21, 2011.

Beaudry, P. and Lewis, E., "Do Male-Female Wage Differentials Reflect Differences in the Return to Skill? Cross-City Evidence from 1980–2000," *American Economic Journal: Applied Economics* 6, 178–94, 2014.

Bloem, A. M., P. Cotterell, and T. Gigantes, "National Accounts in Transition Countries: Balancing the Biases?," *Review of Income and Wealth*, 44, 1–24, 1998.

Bound, J., C. Brown, and N. Mathiowetz, "Measurement Error in Survey Data," in J. J. Heckman and E. Leamer (Eds), *Handbook of Econometrics.* Elsevier, 2001.

Clemenceau, A. and J.-M. Museux, "EU-SILC (Community Statistics on Income and Living Conditions: General Presentation of the Instrument)," in Comparative EU Statistics on Income and Living Conditions: Issues and Challenges, Proceedings of the EU-SILC conference (Helsinki, 6–8 November 2006), 2007.

Clements, B. J., S. Gupta, S. Tareq, and I. Karpowicz, "Evaluating Government Employment and Compensation," IMF Technical Notes and Manuals No. 2010/15, 2010.

Gaetano, D. A., "Wage Spillovers across Sectors in Eastern Europe," *MPRA Paper*. Department of Applied Economics II, Universidad de Valencia, 2011.

Genre, V., K. Kohn, and D. Momferatou, "Understanding Inter-Industry Wage Structures in the Euro Area," *Applied Economics*, 43, 1299–313, 2011.

Gittleman, M., and E. N. Wolff, "International Comparisons of Interindustry Wage Differentials," *Review of Income and Wealth*, 39, 295–312, 1993.

Hartwig, J., "Testing the Baumol-Nordhaus Model with Eu Klems Data," *Review of Income and Wealth*, 57, 471–89, 2011.

Heller, P. S., and A. A. Tait, "Government Employment and Pay: Some International Comparisons," International Monetary Fund Occasional Paper No. 24, 1984.

Hotchkiss, J. L. and Pitts, M. M., "The Role of Labor Market Intermittency in Explaining Gender Wage Differentials," *American Economic Review*, 97, 417–21, 2007.

Hua, L. A., "Assessment of Statistical Data Quality: The Example of the Wages Part in the Ilo Yearbook of Labor Statistics," Conference on Data Quality for International Organizations, Rome, Italy, 7 and 8 July, 2008.

Klitgaard, R., "Incentive Myopia," *World Development*, 17, 447–59, 1989.

Lamo, A., J. J. Pérez, and L. Schuknecht, "Public or Private Sector Wage Leadership? An International Perspective," *The Scandinavian Journal of Economics*, 114, 228–44, 2012.

La Porta, R., F. Lopez-De-Silanes, A. Shleifer, and R. Vishny, "The Quality of Government," *Journal of Law Economics and Organization*, 15, 222–79, 1999.

United Nations, Commission of the European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, and World Bank, 2008 System of National Accounts, United Nations, NY, 2009 (http://unstats.un.org/unsd/nationalaccount/sna2008.asp).

Schiavo-Campo, S., G. De Tommaso, and A. Mukherjee, "An International Statistical Survey of Government Employment and Wages," World Bank Policy Research Working Paper No. 1806, 1997.

Van Rijckeghem, C. and B. Weder, "Bureaucratic Corruption and the Rate of Temptation: Do Wages in the Civil Service Affect Corruption, and by How Much?," *Journal of Development Economics*, 65, 307–31, 2001.

Appenidx A. The Structure of the ISIC 3 Classification of Industries

Industry	Definition	2 Digits	3 Digits	4 Digits
(1)	(2)	(3)	(4)	(5)
A	Agriculture, hunting and forestry	01-02	011-015	0111-0150
B	Fishing	05	050	0500
C	Mining and quarrying	10-14	101-142	1010-1429
D	Manufacturing	15-37	151-372	1511-3720
E	Electricity, gas and water supply	40-41	401-410	4010-4100
F	Construction	45	451-455	4510-4550
G	Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods	50-52	501-526	5010-5260
H	Hotels and restaurants	55	551-552	5510-5520
I	Transport, storage and communications	60-64	601-642	6010-6420
J	Financial intermediation	65-67	651-672	6511-6720
K	Real estate, renting and business activities	70-74	701-749	7010-7499
L	Public administration and defense; compulsory social security	75	751-753	7510-7530
M	Education	80	801-809	8010-8090
N	Health and social work	85	851-853	8511-8532
O	Other community, social and personal service activities	90-93	900-930	9000-9309
P	Private households with employed persons	95	950	9500
Q	Extra-territorial organizations and bodies	99	990	9900

Appendix B. The Structure of ISIC 4 Classification of Industries

Industry	Definition	2 Digits	3 Digits	4 Digits
A	Agriculture, forestry and fishing	01-03	011-032	0111-0322
B	Mining and quarrying	05-09	051-099	0510-0990
C	Manufacturing	10-33	101-332	1010-3320
D	Electricity, gas, steam and air conditioning supply	35	351-353	3510-3530
E	Water supply; sewerage, waste management and remediation activities	36-39	360-390	3600-3900
F	Construction	41-43	410-439	4100-4390
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	45-47	451-479	4510-4799
H	Transportation and storage	49-53	491-532	4911-5320
I	Accommodation and food service activities	55-56	551-563	5510-5630
J	Information and communication	58-63	581-639	5811-6399
K	Financial and insurance activities	64-66	641-663	6411-6630
L	Real estate activities	68	681-682	6810-6820
M	Professional, scientific and technical activities	69-75	691-750	6910-7500
N	Administrative and support service activities	77-82	771-829	7710-8299
O	Public administration and defense; compulsory social security	84	841-843	8411-8430
P	Education	85	851-855	8510-8550
Q	Human health and social work activities	86-88	861-889	8610-8890
R	Arts, entertainment and recreation	90-93	900-932	9000-9329
S	Other service activities	94-96	941-960	9411-9609
T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	97-98	970-982	9700-9820
U	Activities of extraterritorial organizations and bodies	99	990	9900

Appendix C. The Correspondence Between SEDLAC Sectors and ISIC Industries

ISIC	Primary activities	Industry low tech	Industry high tech	Constructio	Commerce	Utilities & transportation	Skilled services	Public administration	Education & Health	Domestic servants
ISIC 3	A - Agriculture, forestry and B - Fishing C- Mining and quarrying	D - Manufacturing		F - Construction	G - Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods H - Hotels and restaurants	E - Electricity, gas and water supply I - Transport, storage and communications	J - Financial intermediation K - Real estate, renting and business activities O - Other community, social and personal service activities	L - Public administration and defence; compulsory social security Q - Extraterritorial organizatios and bodies	M -Education N - Health and social work	P - Private households with employed persons

Manufacturing (industry D) wages from SEDLAC are the average of the wages of the low tech and high tech sectors, weighted by the labor share of the two sectors in the sample. Suppose that a sample of N manufacturing workers is surveyed, of which N_l are in the low tech sector and N_h in the high tech sector. The corresponding average wages are W_l and W_h. The average wages for the manufacturing industry are then obtained as

$$W_m = \frac{N_l}{N_l + N_h} W_l + \frac{N_h}{N_l + N_h} W_h$$

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix D: List of countries and data sources included in the database.