Darmstadt
Discussion Papers
in ECONOMICS

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Determinants of ICT Infrastructure: A Cross-Country Statistical Analysis

Jens J. Krüger, Mathias Rhiel

Nr. 228

Arbeitspapiere der Volkswirtschaftlichen Fachgebiete der TU Darmstadt
ISSN: 1438-2733

ISSN 1438-2733

Printed in Germany
Technische Universität Darmstadt
Department of Law and Economics
D – 64283 Darmstadt
Germany
www.wi.tu-darmstadt.de

Determinants of ICT Infrastructure: A Cross-Country Statistical Analysis

October 2016

by

Jens J. Krüger

Darmstadt University of Technology
Department of Law and Economics
Hochschulstraße 1, D-64289 Darmstadt, Germany
tel.: +49 6151 1657281, fax: +49 6151 1657282
e-mail: krueger@vwl.tu-darmstadt.de

and

Mathias Rhiel

Darmstadt University of Technology
Department of Law and Economics
Hochschulstraße 1, D-64289 Darmstadt, Germany
tel.: +49 6151 1657284, fax: +49 6151 1657282
e-mail: rhiel@vwl.tu-darmstadt.de

Abstract

We investigate economic and institutional determinants of ICT infrastructure for a broad cross section of more than 100 countries. The ICT variable is constructed from a principal components analysis. The explanatory variables are selected by variants of the Lasso estimator from the machine learning literature. In addition to least squares, we also apply robust and semiparametric regression estimators. The results show that the regressions are able to explain ICT infrastructure very well. Major determinants identified are real income per capita, the availability of electricity, the extent of urbanization and indicators for the quality of the institutional environment. We also find evidence of conditional convergence of the ICT infrastructure across countries.

JEL classification: O11, O33, L96, C52, C14

Keywords: determinants of ICT infrastructure, global digital divide, variable selection, machine learning, robust and semiparametric regression

1 Introduction

Information and communication technologies (ICT) as determinants of macroeconomic growth will be of increasing importance in the coming decades. Empirical evidence for the growth effects of ICT are provided, for instance, by Jorgenson and Stiroh (1999), O'Mahony and Vecchi (2005) and van Ark et al. (2008). Therefore it is interesting also for policy purposes what determines ICT infrastructure. Studies pursuing this line across countries are, inter alia, Hargittai (1999), Caselli and Coleman (2001), Kiiski and Pojola (2002) and Chinn and Fairlie (2007). The estimates reported in the received literature are based on different theoretical approaches and therefore lead to different sets of explanatory variables. Thus, depending on the theoretical stance some variables are considered in the some studies whereas others are entirely neglected. What is lacking is a comprehensive approach to consider a broad set of candidate explanatory variables simultaneously and to use modern methods for model selection to determine the optimal set of explanatory variables.

In this paper we pursue such an approach based on variable selection methods originating from machine learning research. These methods, the so-called Lasso and its variants, are based on regularization instead of significance tests. Their application leads to parsimonious regression specifications using the most relevant explanatory variables and reaching a high degree of fit. Although no economic theory is involved in the variable selection procedure, theory is of course involved in assembling the set of candidate variables and in the interpretation of the results. We investigate economic and institutional determinants of ICT infrastructure for a broad cross section of more than 100 countries at very different stages of development. To reach a comprehensive measure of ICT infrastructure, several indicators for the period 2002-2012 are combined by a principal components analysis.

The empirical approach followed in this paper relies on three distinctive features which are novel to the literature. First, principal components analysis is used to construct the dependent variable for ICT infrastructure from bundling several indicators in an optimal way. Second, methods from the machine learning literature (the so-called Lasso and several of its more advanced variants) are used to select the relevant explanatory variables from a broad set of candidates. Third, in addition to common least squares regression, recent methods for robust regression estimation and semiparametric regression are used to validate the results against the influence of outliers in the data and to uncover nonlinear effects of the explanatory variables, respectively.

Overall, our findings show that the variables selected by our approach can explain the bulk of cross-country differences in the state of ICT infrastructure. Major determinants identified are income per capita, the availability of electricity, the extent of urbanization and indicators for the quality of the institutional environment. These findings are sustained when using the robust regression estimator. Splitting the whole sample period 2002-2012 into subperiods reveals that conditional convergence dynamics are operating here also. Moreover, for some explanatory variables nonlinear effects can be established.

The analysis proceeds by providing an extensive literature review in section 2, identifying a wide range of relevant explanatory variables. This is followed by the description of the database and the specific construction of the dependent variable from a principal components analysis in section 3. The variable selection approach and the route taken for the empirical analysis are outlined in section 4. The results are presented and discussed in section 5. Section 6 provides concluding remarks.

2 Literature Review

In this section we review the existing literature on the determinants of ICT infrastructure. This serves to demonstrate the state of research and to identify those determinants previously scrutinized. Basically, many empirical studies have been conducted both at the micro and the macro level to discover the determinants of ICT in general. The studies on the micro level examine the factors influencing the firm's investment behavior in ICT. At the macro level, the literature discusses cross-country differences in the adoption of information technology, mostly in context of the *global digital divide* between advanced and less developed countries.[1] In our paper we focus on the factors explaining the ICT infrastructure at the macro level.

Most of the studies share the same approach of first making theoretical assertions about factors influencing ICT, followed by identifying appropriate indicators for these factors. In a second step, the dependent ICT variable is regressed on the identified explanatory variables. The studies basically differ in the specification of the dependent variable and therefore in the concretization of the research object ICT. Common dependent variables used in the literature are ICT expenditure, number of internet users, adoption of internet by employees or ICT imports. Furthermore, the individual investigations differ regarding time coverage and country sample.

[1] The term of the (global) digital divide is extensively discussed in Norris (2001) and Hargittai (2003).

As a result it is not surprising, that quite diverse findings are reported in the literature. One the one hand, a set of common ICT-explaining variables is used in the studies. On the other hand, another group of variables is mentioned in the literature only in single occasions. Generally, the variables can be classified in the following categories, ordered by decreasing importance in the literature: the economic wealth and structure of the countries, human capital, regulations, demographic factors and geographical/territorial factors. We now review the results of this literature structured along these categories.

Economic Wealth and Structure

Per capita income is the main and mostly used determinant of ICT (see e.g. Hargittai 1999, Kiiski and Pojola 2002, Norris 2001, Beilock and Dimitrova 2003). In the previous literature it has been found that countries whose citizens are better off economically tend to have more ICT (see e.g. Hargittai 1999, Beilock and Dimitrova 2003). The underlying assumption is that countries with higher per capita income invest more in R&D and therefore are more able to discover and better in adopting ICT (Baliamoune-Lutz 2003). So, per capita income influences ICT indirectly. Next to the education, income is an important determinant of computer ownership and internet use (OECD 2001). Per capita income is found to be positively and significantly related to ICT adoption by Caselli and Coleman (2001), Guillén and Suárez (2001), Kiiski and Pojola (2002), Baliamoune-Lutz (2003), Pohjola (2003), Chinn and Fairlie (2007), Wunnava and Leiter (2009). In contrast, Dasgupta et al. (2001) find the relationship as not significant.

Next to the level of per capita income, economic conditions are also characterized by income equality within a country which "may have a negative effect on ICT diffusion because fewer people will be able to afford to pay for ICT products and services" (Wunnava and Leiter 2009, p. 418). Hargiatti (1999) examined (among other variables) the impact of income equality on internet connectivity among OECD countries, but did not find a significant relation.

The sectoral composition of the economy has also been considered in the literature. As the underlying idea the share of manufacturing and/or service sector are supposed to positively affect investment rates in ICT. Caselli and Coleman (2001) found no evidence supporting this assertion. Despite of this, they found evidence for an inverse relationship with the share of the agricultural sector. A positive effect of employment in the service sector (as percent of total) and negative in public sector has been found by Gust and Marquez (2004).

Human Capital

Next to the differences in economic wealth of countries, human capital is frequently addressed in the literature. The basic idea for considering human capital as a determinant of ICT is that skilled and educated workers are more capable of learning how to use new technologies. Especially academic institutions play an essential role in adopting new technology (Guerrieri et al. 2011). While schools were among the first to introduce young people to ICT, these technologies provide the basis for research and education today and promote their adoption also in this way.

From the theoretical point of view, human capital seems to be one of the most essential factors which are positively influencing ICT adoption. Empirically, however, most of the authors found no clear evidence for this hypothesis. Wunnava and Leiter (2009) found significantly positive effects of tertiary enrollment on internet diffusion. Baliamoune-Lutz (2003) uses the education index from the UNDP Human Development Report as a variable for human capital, finding a positive effect on the diffusion of mobile telephones, but no effect on the diffusion on internet hosts, internet users or personal computers. Crenshaw and Robinson (2006) used tertiary education enrollments and Chinn and Fairlie (2007) chose the years of schooling as a determinant for human capital. Both also find mixed evidence for the role of human capital as a determinant of ICT.

Gust and Marquez (2004), using years of schooling as determinant for ICT expenditures, and Hargiatti (1999), relying on the education index from the UNDP Human Development Report, find a significant and positive effect on OECD countries. Kiiski and Pohjola (2002), however, came to a different result using the average years of schooling for the population over age 15, obtained from Barro and Lee (see Barro and Lee (2000)), to discover the effect on internet diffusion in the OECD.

Thus, education does not seem to explain global differences in ICT robustly. In contrast to studies on the macro level, a positive relationship between ICT and employee qualification can be found on the micro level at Bayo-Mariones and Lera-Lopez (2007), as well as Haller and Traistaru-Siedschlag (2007).

Intuitively, a low level of education obstructs both the accessibility and distribution of ICT. A form of a particularly low education level is illiteracy. Literacy is required because of the text-based technologies of application software, world wide web and e-mail. However, the application range has been expanded in the last years. For example, video and voice communication applications do not necessarily need a higher level of literacy. The effect of literacy has been examined in studies of Baliamoune-Lutz (2003) as well as Chinn and Fairlie (2007).

Baliamoune-Lutz (2003) uses literacy rates of adults as an indicator of the initial level of education. She found no significant effect of literacy on ICT. Chinn and Fairlie (2007) neither find a significant effect of illiteracy rates on computer penetration rate nor internet penetration rate. In general, data on literacy[2] are limited. Behrman and Rosenzweig (1994) pointed out a major problem in using literacy data for cross-country comparisons lies in differences of its definition. As another problematic issue, actual data on which literacy rates are based are often sparse and dated. Beside this critique on the definition and calculation, Barro and Lee (2013) found literacy rates not adequately measure the aggregate stock of human capital. Although frequently used, literacy rates do not seem to be an appropriate variable to capture human capital.

Besides general education and literacy, the knowledge of the English language is an important aspect of ICT usage. Because English is the most important language in the ICT domain, most of the software, internet sites and internet-supported communication is shaped in that language. In connection with higher education, most of the scientific and academic work is taught, written and published in English. Caselli and Coleman (2001) are not able to find a significant effect for the English language skills of the population on computer imports per worker. Kiiski and Pohjola (2002) measure English skills by the percentage of pupils in secondary education learning English from the European Commission. The lack of data reduces the number of observations to 17 countries, for which they significantly find a negative sign in the regression. Guillén and Suárez (2001) include a dummy variable to identify countries in which English is an official language or the most widely spoken language.[3] They find English to be positively related to the worldwide number of internet users and hosts.

In summary, it can be concluded that human capital theoretically is one of the most plausible factors for explaining ICT. However, the empirical evidence is rather mixed and can not robustly identify a relation in several studies using various indicators for human capital.

Regulation

The impact of regulation on ICT adoption is a widely discussed topic in the literature. The basic argument is that all kinds of regulations or constraints hinder individuals in acting optimal (Guerrieri et al. 2011). The regulation aspect is particularly relevant in interaction with the prosperity level of nation. The idea is that "richer countries have well-developed market economies and well-established legal systems, and as a result are able and willing to invest more in research and development and innovation" (Wunnava and Leiter (2009), p. 416).

Popular variables measuring the extent of regulation from the literature are indexes for property rights and civil liberties, used by Caselli and Coleman (2001), Norris (2001), Baliamoune-Lutz (2003) or Crenshaw and Robinson (2006). The results show mixed evidence. Caselli and Coleman (2001) find a positive influence of property rights on the computer imports per worker, but only for a specific set of 45 countries. Baliamoune-Lutz (2003) find property rights explaining the diffusion of mobile telephones and internet hosts significantly. She finds neither an effect on the diffusion of internet users nor on the diffusion of personal computers. In her examination civil liberties only have a significant and positive relation on the diffusion of internet hosts. Crenshaw and Robinson (2006) find property rights explaining the diffusion of internet hosts significantly.

Dasgupta et al. (2001) included the aspect of competition policy in their analysis. They argue that measures of competition policy "affect both the supply of internet services and the intensity of their use by local firms" (Dasgupta et al. (2005, p. 3)). This idea can be transferred from internet services to the entire ICT. As a proxy for government competition policy Dasgupta et al. use the variable 'Government Inhibition of Competition in the Private Sector' from World Bank database. This variable, varying from 1 (most inhibition of a competitive private sector) to 6 (least inhibition), indicates whether the country inhibits a "competitive private sector, either through direct regulation or by reserving significant economic activities for state-controlled entities" (Dasgupta et al. (2005, p. 3)). The authors find evidence for their hypothesis that a low level of inhibition has a significant and positive effect on the diffusion of internet and mobile phone subscribers.

Another aspect examined in the literature is the market structure of the telecommunications sector. Basic idea here is that competition in the telecommunications market leads to reduced prices for access and use. The results are again ambiguous. Hargiatti (1999) found a negative influence of a telecommunications monopoly on the internet connectivity in industrialized countries. The evidence of such a negative influence could not be confirmed by Kiiski and Pohjola (2002) and only partially by Guillén and Suárez (2001, 2005).

Gust and Marquez (2004) establish a negative influence of regulation in the labour market on ICT spending. They use three indexes: an index of employment protection legislation (from the OECD), an index of regulatory burdens on startups (World Economic Forum) and an index on overall regulatory burdens (World Economic

[2] Both, the adult literacy rate of the population over age 15 and illiterate population over age 15.

[3] They also include a dummy variable for Scandinavian countries, in which an unusually large percentage of the population knows English as a second language.

Forum). All three indexes had a negative and significant influence on the ICT expenditures of 13 industrialized countries during the period 1992-1999.

As can be seen, various aspects and variables exist concerning the subject of regulation. In conclusion, the subject of regulation is important for an explanation of the global digital divide. However, a definite variable capturing the degree of regulation could not be identified so far and is not in reach because of the multi-faceted nature and great diversity of regulatory measures.

Demographic Factors
As another aspect, demographic factors have received attention in the literature. The hypothesis states that the age structure and the size of the urban population explain ICT. The underlying idea is that young people and the urban population in general tend to use more ICT because of network economies and firms being mostly located in cities or in their neighborhood. Concerning the age structure, no effort has been made to determine empirical evidence. Chinn and Fairlie (2007), however, suggest, that "the global digital divide would be even larger if developing countries had an age composition that was more similar to the United States" (Chinn and Fairlie 2007, p.18).

The share of cities in the production of national gross domestic product ranges from an average of 55% in the developing world to 85% in developed countries (Crenshaw and Robinson 2006). Therefore it would be plausible that a higher degree of urbanization positively influences ICT diffusion. Both Dasgupta et al. (2001) and Crenshaw and Robinson (2006) find a positive effect of urban population. Chinn and Fairlie (2007), however, get a negative effect.

Geographical / Regional Factors
To control for geographical and regional factors, several authors include respective dummy variables. The usage of these variables uncovers the influence of explanatory variables on ICT for a specific group of countries. Kiiski and Pohjola (2002) include dummy variables for "nordic" countries, "southern" countries as well as for Mexico and Turkey. Beilock and Dimitrova (2003) divide the world into six regions and test for differences in internet usage rates. The regions considered are highly developed nations, Latin America, formerly socialist nations, Middle East/North Africa, rest of Asia and Sub-Saharan Africa. Only the parameter estimate associated with the rest of Asia is found significant with a positive sign.

Interim conclusion from the literature review

In this section we reviewed the relevant literature on determinants of ICT infrastructure on the macro level. During the last 15 years, several attempts have been made to explain the *global digital divide*. As to be expected, the results are quite diverse. Some variables, like the GDP per capita, were unambiguously identified as a major determinant of ICT. A variety of variables are mentioned in the literature only once. Surprisingly, some variables or groups of variables have shown no significant influence despite their clear theoretical relevance. Even though human capital is one of the most featured factors in theory, the empirical evidence could not be consistently established in several studies using various indicators of human capital.

Taken together, the bottom line is that the question of what the *global digital divide* explains is not conclusively answered yet. In the following, we will use the insights gained from the literature reviewed above to build up a encompassing database of candidate variables which are potentially relevant for explaining ICT infrastructure. For these candidate variables we entertain a specific variable selection approach to find the variables which are most relevant for explaining our indicator of ICT infrastructure. Next, we turn to the construction of this indicator and the description of the database in general.

3 Data

The data used for forming the dependent variable and constituing the set of candidate explanatory variables are assembled from various sources. We will first focus on the dependent variable for ICT infrastructure which we construct as the first principal component from a principal components analysis. The variables serving as candidates for the model selection procedure used in this paper are described subsequently.

Construction of the ICT Infrastructure Variable

In both publicly accessible databases and the literature a single encompassing variable describing all aspects of ICT infrastructure can not be identified. Previous studies often use variables which describe single isolated aspects only.[4] Examples for variables describing isolated aspects of ICT infrastructure accessible in public databases are:

- the number of telephone lines (per 100 people),
- the number of internet users (per 100 people),
- the number of broadband internet subscribers (per 100 people),
- the number of mobile cell subscribers (per 100 people),
- the number of PCs (per 100 people).[5]

Almost all of the data are from the World Telecommunication/ICT Development Report of the International Telecommunication Union (ITU), available for the years from 2000 onwards and can be downloaded from World Bank database. The variables are highly correlated with each other (see appendix table 5).[6] Using a principal component analysis (PCA), the strongly correlated variables can be reduced to a single meaningful variable. Table 1 summarized the results of the PCA using averaged data for the years 2002-2012.[7]

Table 1: PCA output

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Telephone Lines	0.284	-0.297	-0.517	0.738	-0.140
Internet Users	0.388	-0.414	-0.444	-0.665	-0.198
Broadband Internet Subscribers	0.095	-0.192	-0.133	-0.024	0.967
Mobile Cell Subscribers	0.824	0.541	0.159	0.022	0.049
PCs	0.285	-0.640	0.702	0.113	-0.056
Standard derivation	86.277	20.734	8.149	6.090	2.594
Proportion of Variance	0.932	0.054	0.008	0.005	0.001
Cumulative Proportion	0.932	0.986	0.994	0.999	1.000

Note: All Variables are in values per 100 people and averaged over the years of 2002-2012.

As we can see in Table 1, the first principal component describes 93.2% of the entire variance. All variables load on the first principal component and all loadings have a positive sign. The other components add only a small amount to the explained variance. As a result of the PCA, the five variables of ICT infrastructure can be collapsed to a single variable that comprises most of the information. It is evident that we can not be sure whether our variable mainly reflects the infrastructure, the equipment or the usage of ICT. It is also clear that infrastructure, equipment and usage mutually rely on each other. Therefore, we might have used these (and related) terms interchangeably, but decided to stick to the term ICT infrastructure in the following discussion.

The boxplot in figure 1 illustrates the distribution of this ICT infrastructure variable, constructed for a total of 178 countries as the first principal component. The values of a few selected countries (right side of the boxplot)

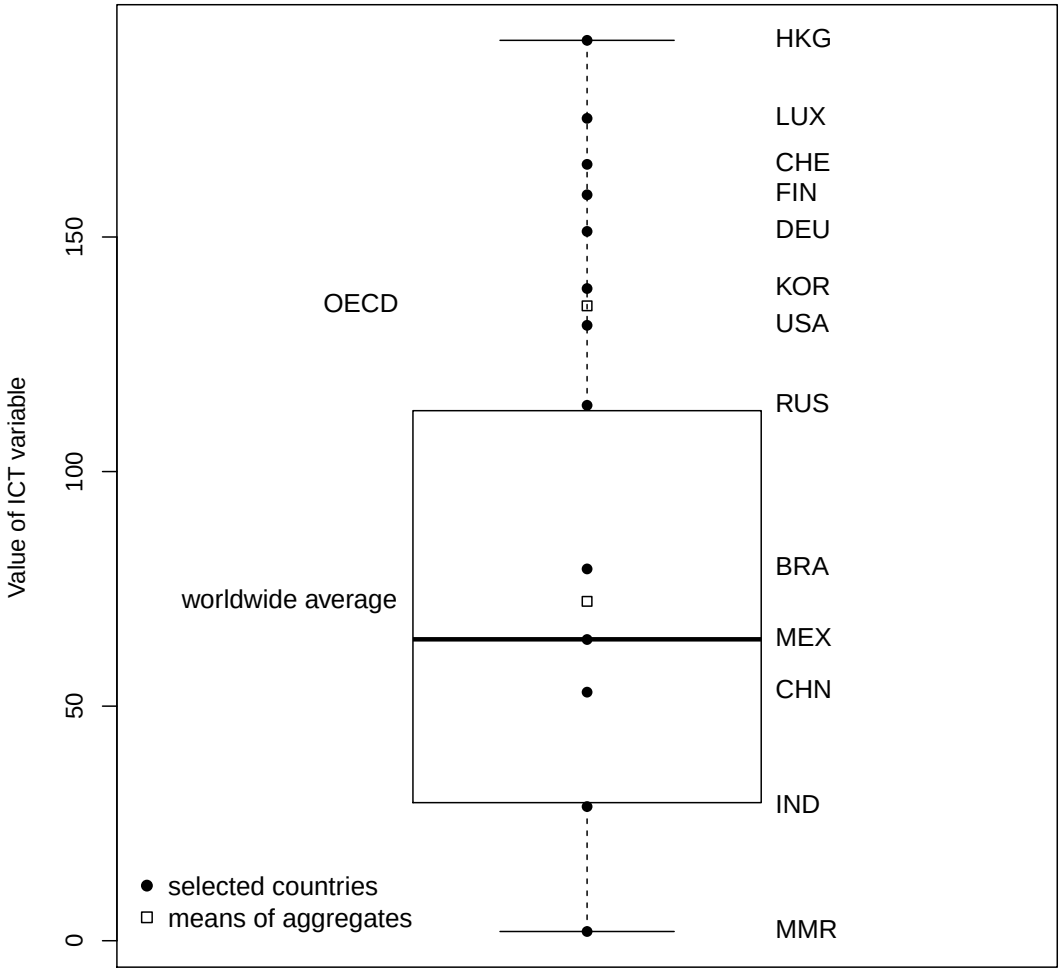
[4] The number of internet hosts was used by Guillén and Suárez (2001) as well as Kiiski and Pohjola (2002). Guillén and Suárez (2001) also used the number of internet users to explain. The number of PCs was explained by Chinn and Fairlie (2007). Wunnava and Leiter (2009) investigated the internet diffusion rates.

[5] The *number of internet users (per 100 people)* captures the individuals who have used the internet via a computer, mobile phone, personal digital assistant, games machine, digital TV or other. With the *number of broadband internet subscribers (per 100 people)*, the number of subscribers with a digital subscriber line, cable modem, or other high-speed technology per 100 people is counted. The *number of mobile cell subscribers (per 100 people)* is the number of subscriptions to a public mobile telephone service using cellular technology, which provide access to the public switched telephone network. Included are postpaid as well as prepaid subscriptions. The *number of PCs (per 100 people)* is the estimate of the number of personal computers of the United Nations. For further description and definition of these data, see homepage of the International Telecommunication Union: http://www.itu.int/en/ITU-D/Statistics/Pages/publications/handbook.aspx.

[6] For example, users connect to the internet either via telephone line or via wireless and mobile technology. For this reason, the fraction of internet users overlaps with the fraction of telephone lines on the one hand and the mobile cell subscribers on the other hand. We can find a similar overlap between the fraction of internet users and mobile cell subscribers with the fraction of broadband internet subscribers. This overlap is of course reflected by the high correlations between the single aspects of IT infrastructure.

[7] To get a more robust explanatory variable, we average the input variables over a period of ten years. The latest data are from 2012. Before 2002, values for the number of broadband internet subscribers (per 100 people) are rarely available because of the relatively new broadband technology.

Figure 1: Boxplot of the ICT infrastructure variable



Note: Distribution of the ICT infrastructure variable.

as well as the averaged values of all countries and OECD in particular (left side of the boxplot) are also indicated in the boxplot. The variable values range from approximately 2 (Myanmar) to a value slightly above 191 (Hong Kong). This implies huge cross-country differences in the stage of ICT infrastructure. The value of Mexico lies close to the median of 64.2, the value of Columbia and Tunisia near the mean of about 72. The box, with the 1st and 3rd quantiles as its lower and upper margins, has a comparably low position which means that 75% of the countries have a value of ICT infrastructure below 112 while a few countries show fairly high values. These are either countries which are mere cities (Hong Kong, Luxembourg) or small advanced countries (Switzerland, Finland). While the worldwide mean value is in the middle of the box, the mean value of the OECD countries lies in the upper whisker between the positions of Korea and the US.[8]

In figure 2 the global distribution of ICT infrastructure is plotted on a world map. Higher values of ICT infrastructure are represented by darker areas. These can be observed in North America (mean value of ICT infrastructure 127.90) and Europe (121.77). South, East (30.33) and West Africa (30.83) as well as Central Asia (34.37) present lower values.

[8] The position of Germany (DEU) appears to be high on the scale, especially compared to the US. Inspecting the input variables for the PCA closer for the subsample of OECD countries reveals that Germany is in no way exceptional with respect to any single variable. It should also be recalled that all variables are expressed on a per capita basis to avoid undue influence of country size.

Figure 2: Worldmap of ICT infrastructure variable($\ln IT$)

Note: Global distribution of the ICT infrastructure variable. Countries with missing data represented by white color.

We use the (log transformed) indicator for ICT infrastructure explained previously, generated as an average of the respective variable values for the period 2002-2012 ($\ln IT$), as well as an average for the subperiod 2008-2012 ($\ln IT_2$). This second choice for the dependent variable allows to investigate the degree of persistence by allowing $\ln IT_1$ (the average variable values for the subperiod 2002-2006) as an explanatory variable.

Explanatory Variables

As described in the literature review above, many variables appear to be used to describe the global differences in ICT or ICT infrastructure. In addition to the variables addressed in the literature, some other variables seem to be potentially relevant. Since the variable selection in the received literature sometimes seems to be arbitrary we pursue a different approach. This approach consists of compiling a large database of potentially relevant candidate variables and use a specific statistical approach (the so-called Lasso explained below) to select the relevant explanatory variables from this pool.

All collected data are freely available on the internet. They come from the World Bank[9], the Quality of Government Institute[10], the Barro-Lee dataset[11], the Heritage Foundation[12], the Penn World Table version 8[13], the International Monetary Fund[14], the database of the United Nations[15] and the dataset of Sala-i-Martin, Doppelhofer and Miller[16]. We averaged all variables over the years of 1980 to 2000 as far as possible and mostly transformed them by taking their natural logarithms. As far as values for the years 1980 to 2010 were

[9] http://databank.worldbank.org/data.
[10] http://qog.pol.gu.se/data/datadownloads.
[11] http://www.barrolee.com.
[12] http://www.heritage.org.
[13] http://www.rug.nl/research/ggdc/data/pwt.
[14] http://data.imf.org.
[15] http://data.un.org/DataMartInfo.aspx.
[16] The dataset was assembled for the paper of Sala-i-Martin, Doppelhofer and Miller (2004). It can be downloaded from https://www.aeaweb.org/aer/data/sept04_bace_data.zip.

available, we also calculated the average growth rate (in logged differences) and the standard deviation of the annual growth rates. In addition to these variables, we also include dummy variables to control for geographical localization[17] and the development stage[18] of a country. All variables, their data source and literature references are listed in table 6 of the appendix.

As a result we obtain a dataset of 72 different variables for a total of 178 countries.[19] Expressed in the categories of the variable classification (see section above), the dataset contains 21 variables for national economic wealth and structure, 12 variables representing human capital, 23 variables measuring the extend of regulations, 3 demographic and 9 geographical/regional variables.[20] These 72 variables are supplemented by their transformations[21] to reach a total of 148 candidate explanatory variables.

The method of variable selection we pursue in this work, requires a dataset with complete observations. Possible approaches to meet this requirement are quite drastic: The exclusion of all variables having at least one missing entry would result in only 4 countries left in the dataset. Vice versa, removing of all countries having a missing entry for at least one variable leads to a "dataset" with no countries included at all. The manual sorting-out of those variables not covering a sufficient amount of countries or of those countries with an insufficient number of variables, requires considerable effort and is inevitably subjective. Thus, an automated algorithmic approach is desirable.

To reach a more objective decision, we compute the percentage of available variable values per country. We then exclude all countries below a certain threshold of available variables values. From the remaining countries we exclude This procedure results in a complete dataset. Using a threshold of 0.835 we reach a dataset with 81 variables and 113 countries, containing all OECD countries but also many developing countries. This is the typical sample size also reached in many cross-country growth analyses. Countries and variables of the final dataset are described in table 7 and6, some descriptive statistics of these variables in table of the appendix. Also included in the appendix are the density plots of the three ICT infrastructure variables for the final dataset with 113 countries (see figure 6). The plots show that the two levels variables $\ln IT$ and $\ln IT_2$ have rather similar left-skewed densities. We used this dataset for the subsequent analysis utilizing a bunch of statistical methods which are described in the following section.

4 Method

In this work we intend to select explanatory variables from a large pool of candidate variables by methods originating from machine learning research (see e.g. Murphy (2012) for a recent comprehensive account of this field). The aim of the selection procedure is to find appropriate variables for explaining cross-country differences in ICT infrastructure. These variables are subsequently introduced into a regression analysis. For the estimation we use ordinary least squares with a heteroskedasticity-consistent covariance matrix as well as a robust regression estimator and a semiparametric generalized additive model (GAM) estimator.

As outlined above, several ICT influencing variables have been identified in the literature. The results are diverse and partly contradictory. In our database of candidate explanatory variables we record those variables identified in the literature as well as generally potentially relevant variables. Since there are nearly as much explanatory variables in the database as country observations, the above mentioned methods of variable selection are used to get a parsimonious model with stable and unbiased coefficient estimates. These methods are currently diffusing from the machine learning area into econometrics.[22]

The statistical approach pursued is based on a linear regression model stated for country i out of a cross section of n countries

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i,\ i = 1, ..., n\,, \tag{1}$$

[17] Dummy variables for: East Asia and the Pacific, Europe and Central Asia, Latin America and the Caribbean, South Asia, Sub-Saharan Africa. Data is provided by Barro-Lee and available on the internet at http://www.barrolee.com/.

[18] Dummy variables for: Advanced Economies, developing countries and developed countries, countries of the OCED and countries of the European Union. Data of the first three groups is provided by Barro-Lee and available on the internet at http://www.barrolee.com.

[19] We reduced the initial dataset from 255 countries to 178 countries for which data for the dependent IT infrastructure variable are available. A list of the countries included in the initial as well as in the reduced dataset is shown in table **??** of the appendix.

[20] 4 further variables do not fit to the variable classification.

[21] As far as meaningful, the variables are logarithmized. In addition, we also calculated the growth rate and standard deviation of the annual growth rates (in logged differences) for a specific variable if this is sensible.

[22] For more on machine learning methods in an econometric context see inter alia Bajari et al. (2015a,b), Belloni et al. (2012), Doornik and Henry (2015), Kleinberg et al. (2015), Schneider and Wagner (2011) and Varian (2014).

where y_i denotes the dependent variable (an indicator of ICT infrastructure in our case), $\boldsymbol{x}_i$ is the k-vector of explanatory variables (including a constant) and u_i is the usual error term.

The machine learning methods for variable selection rely on regularization which amounts to add a penalty term to the least squares target function. The motivation is that larger coefficient estimates tend to induce higher variability in the least squares fit. Whereas the OLS estimator is unbiased under the classical assumptions, regularization tolerates some bias in order to reduce the variance. The Lasso (least absolute shrinkage and selection operator) regression, proposed by Tibshirani (1996), performs a selection of variables by introducing a specific penalty term weighted by a factor $\lambda > 0$. This term penalizes the magnitude of the regression coefficients in the vector $\boldsymbol{\beta}$ and thereby leads to a complete removal of some variables from the set of candidate explanatory variables.

The Lasso estimator minimizes the target function

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda \cdot \sum_{j=1}^{k}|\beta_j|\,, \tag{2}$$

where the usual least squares target function is augmented by the regularization term serving to penalize large magnitudes of the regression coefficients. The amount of regularization is controlled by the parameter λ which may be chosen by cross-validation methods or information criteria.

The specific form of the regularization term used here causes some coefficients to be forced exactly to zero and thus excludes the associated explanatory variables completely. Those variables increase the penalty term by their regression coefficients but are not able to reduce the residual sum of squares by a substantive amount. This has the beneficial side effect of also reducing multicollinearity. Multicollinearity is usually a problem in large cross-country data sets, because many variables reflect the general state of development of the countries and thus are highly correlated. The Lasso tends to select only those explanatory variables with mild multicollinearlity of each other (see Bajari et al. (2015a)).

A refinement of the basic idea is the adaptive Lasso proposed by Zou (2006), augmenting the penalty term by weight factors, i.e.

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda \cdot \sum_{j=1}^{k} w_j|\beta_j|. \tag{3}$$

In the modified formula w_j denotes the weight factor of the j-th regression coefficient. In this work, we rely on the standard error adjusted adaptive Lasso (SEA-Lasso) proposed by Qiang and Yang (2013). For the SEA-Lasso the weights are defined by $w_j = \hat{\sigma}_j/|\hat{\beta}_j|$, where $\hat{\beta}_j$ are the OLS coefficient estimates and $\hat{\sigma}_j$ the associated standard errors. With this weighting scheme the SEA-Lasso has the advantage of being scale-independent. Moreover, the adaptive variants have the so-called oracle property (see Zou (2006)) as demonstrated by Qian and Yang (2013) for the SEA-Lasso. The oracle property means that asymptotically the adaptive Lasso consistently selects the right variables (those with $\beta_j \neq 0$) and leads to a $\sqrt{n}$-consistent asymptotically normal estimator.

Since we need OLS estimates for forming the weights w_j in the penalty term of the SEA-Lasso target function, we use the so-called Elastic Net before applying the SEA-Lasso for the final variable selection. This procedure is sensible here although we are faced with $n > k$. When k is not much smaller than n and we have considerable collinearity in the data, the OLS estimates would be very unstable and the standard errors tend to be overestimated. In this case the weights could be heavily biased. To deal with this problem, we perform a pre-selection of variables before applying the SEA-Lasso. For this pre-selection we use the Elastic Net (Zou and Hastie (2005)) which combines the basic Lasso with traditional ridge regression (Hoerl and Kennard (1970)). It can be implemented by minimizing the modified target function

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda \cdot \left(\alpha \cdot \sum_{j=1}^{k}|\beta_j| + \frac{1-\alpha}{2} \cdot \sum_{j=1}^{k}\beta_j^2\right)\,, \tag{4}$$

where $\alpha \in [0, 1]$ denotes an additional parameter, controlling the relative importance of the two penalty terms. The parameter α determines whether the penalty term is more akin to the Lasso (in the case of $\alpha = 1$) or more that of a ridge regression ($\alpha = 0$). For $\alpha = \frac{1}{3}$ both penalties are equally weighted. With a pre-selection of variables by means of this procedure, we can reduce biases in the weights used for the SEA-Lasso.

Our statistical approach for variable selection and coefficient estimation can be summarized by the following three-stage procedure:

1. Application of the Elastic Net for the pre-selection of variables. The penalty weight λ is choosen by cross-validation (actually by using 10 randomly assigned folds, repeated 100 times and averaged).
2. Final selection from the pre-selected variables from the previous stage through the application of SEA-Lasso. The penalty weight λ is here chosen by the Bayesian information criterion.[23]
3. Re-estimation of the regression with the selected variables by OLS, the robust Koller-Stahel estimator and GAM regression.

The final re-estimation stage is motivated as follows. All variants of Lasso select a subset of variables and shrink all coefficients towards zero by penalizing their absolute values. As described, regularization lowers the variance with tolerance of some bias. To reduce this bias, we re-estimate the final specification by least squares and robust regression analysis. Actually, we use the OLS estimator with heteroskedasticity-robust standard errors with the correction of MacKinnon and White (1985). We also use the robust regression estimator by Koller and Stahel (2011) which combines the advantage of a high breakdown point[24] with high estimation efficiency.

To uncover nonlinear effects and for validating linear relations we use an additional semiparametric GAM estimator. The GAM is formally stated as

$$y_i = s_1(x_{i1}) + ... + s_h(x_{ih}) + u_i,\ i = 1, ..., n\,, \tag{5}$$

where h denotes the number of selected explanatory variables from the previous stages and the functions $s_j(\cdot)$ are represented by splines. We use Wood's penalized likelihood approach as described in Wood (2001, 2006) for the computation in combination with thin plate regression splines to avoid the choice of knot locations.

The variable selection methods may neglect explanatory variables associated with coefficients of small magnitude which may simply be a consequence of the scaling of the variables (see Chernozhukov et al. (2015, p. 487)). To counteract this tendency, we standardize the explanatory variables for the use in the first two stages. In the subsequent estimation of stage 3, we us the original (not standardized) variables.

As an alternative mode of analysis we apply a bootstrap version of Lasso, the so called bolasso (see Bach (2008)), instead of the first two stages. This variant runs the Lasso for several bootstrap replications of a given sample, on the basis of a residuals bootstrap. This method is proved to be a consistent model selection method under a wider range of conditions than the basic Lasso. We use a soft variant, keeping all variables that are selected in 90 percent of the bootstrap replications. Quite naturally, we find fewer variables using this method. These variables, however, can be viewed as the core predictors that are found to be robustly correlated with the dependent variable in at least 90 percent of the bootstrap replications. In addition, the bolasso can also be used as a device to combat the uniqueness problem of the Lasso in the presence of discrete regressor variables (see Tibshirani (2013)).[25]

5 Results

We now turn to the presentation of the results from the variable selection procedure. This section is divided into three subsections, discussing the results for explaining the three variables introduced above in turn. The guiding idea is to use explanatory variables from a period before the period for which the ICT variable is constructed in order to reduce simultaneity bias.

5.1 Explaining ICT Infrastructure During 2002-2012

The regression results obtained with the three-stage procedure for the variable selection with $\ln IT$ (country means during 2002-2012) as the dependent variable are shown in table 2. The table contains the results of the OLS estimation with the heteroskedasticity-robust standard errors and those of the robust KS regression (reported in parentheses below the regression coefficients are the p-values of the standard t-tests). In addition, we present the results for the semiparametric GAM regression with the nonparametrically entered variables indicated by $s(\cdot)$ (for these variables we report the equivalent degrees of freedom (edf) jointly with the p-values of the F-tests for the joint significance of the spline terms in parentheses). The regressions rely on a total of $n = 113$ observations. In the case of the dummy variables such as region dummies (e.g. EU) we again report the regression coefficients with the p-values of the associated t-tests in parentheses.

[23] For detailed description see Qian and Yang (2013, pp. 298f.).

[24] The breakdown point is defined as the smallest the fraction of contaminated observations in the sample that can lead to an arbitrarily large deviation of the estimator.

[25] All computations are programmed in R using the following packages: *glmnet* and *seaLasso* (for the variable selection), *car*, *lmtest* and *sandwich*)for the least squares regression with the computation of variance inflation factors and the heteroskadasticity-robust standard errors), *robust* (for the robust regressions) and *mgvc* (for the estimation of the GAM). The bootstrap Lasso is implemented in the package *mht*.

Table 2: Regression Results for the Three-Stage Procedure (dependent variable is $\ln IT$)

	OLS	KS		GAM
c	0.183	-0.043	c	4.260
	(0.757)	(0.914)		(0.000)
Elec_m_log	0.250	0.186	s(Elec_m_log)	2.789
	(0.001)	(0.000)		(0.000)
EU	0.101	0.098	EU	0.079
	(0.048)	(0.054)		(0.127)
Europe.and.Central.Asia	0.084	0.077	Europe.and.Central.Asia	0.060
	(0.206)	(0.166)		(0.300)
freedom_corruption_m	0.002	0.001	s(freedom_corruption_m)	1.000
	(0.376)	(0.615)		(0.400)
freedom_corruption_m_log	0.107	0.131	s(freedom_corruption_m_log)	1.000
	(0.240)	(0.098)		(0.339)
gfcf_m_log	0.243	0.199	s(gfcf_m_log)	1.000
	(0.001)	(0.002)		(0.000)
inv_freedom_m	0.003	0.003	s(inv_freedom_m)	1.530
	(0.030)	(0.008)		(0.027)
POP_sd	-6.037	-5.202	s(POP_sd)	2.046
	(0.214)	(0.087)		(0.099)
pyr_m_log	0.059	0.038	s(pyr_m_log)	1.000
	(0.398)	(0.490)		(0.225)
RGDPP_m_log	0.124	0.202	s(RGDPP_m_log)	2.962
	(0.230)	(0.002)		(0.024)
RGDPW_m	0.000	0.000	s(RGDPW_m)	1.000
	(0.398)	(0.732)		(0.559)
South.Asia	-0.411	-0.384	South.Asia	-0.439
	(0.000)	(0.000)		(0.000)
Sub.Saharan.Africa	-0.141	-0.171	Sub.Saharan.Africa	-0.205
	(0.090)	(0.008)		(0.006)
UrbanPop_m_log	0.121	0.117	s(UrbanPop_m_log)	1.000
	(0.166)	(0.019)		(0.024)
R^2	0.948	0.951	R^2	0.955
n	113	113	n	113

Note: Reported are the regression coefficients or the equivalent degrees of freedom (edf) in the case of the spline variables indicated by $s(\cdot)$. Stated in parentheses are p-values of the t-statistics of the F-statistics for the significance of the respective splines.

Considering first the linear regression results in the first three columns of the table we find not all explanatory variables significant here. This is not a contradiction since the variable selection approach relies on regularization and not on significance testing. We find statistically significant coefficient estimates (at 5 percent level of significance) for the explanatory variables access to electricity (Elec_m_log), the European Union dummy (EU, OLS estimator only), gross fixed capital formation (gfcf_m_log), investment freedom (inv_freedom), the South Asian dummy (South.Asia), the Sub Saharan Africa dummy (Sub.Saharan.Africa, KS estimator only), urban population (UrbanPop_m_log, KS estimator only) and expenditure-side real GDP at chained PPPs (in millions 2005 US$) per person (RGDPP_m_log, KS estimator only). Expressed in the categories of the variable classification, the variable selection contains three variables explaining national economic wealth and structure, one variable measuring the extend of regulations, one demographic and two geographical/regional variables.[26] The 'm' in the abbreviation indicates the respective variable as averaged over the years of 1980 to 2000, 'sd' denotes the standard deviation of the respective variable.

In alphabetical order, the variable of access to electricity (Elec_m_log) is the first variable significantly explaining the ICT infrastructure during 2002-2012. This result is not surprising as ICT goods and services need power supply for their operation. The presence of electricity can be seen as an essential prerequisite for ICT infrastructure. The coefficient estimate is an elasticity and its value indicates that ICT infrastructure is inelastic with respect to access to electricity. Surprisingly, we could not find a consideration of this fundamental variable in the literature. Related is the study of Chinn and Fairlie (2006, 2010) using a variable to capture the electric power consumption (kWh per capita) for analyzing cross-country differences in computer and internet penetration. In their study, they find no relationship between per capita electricity use and ICT penetration.

The EU dummy indicates that countries of the European Union on average have higher values of ICT infrastructure compared to the whole country set. In contrast, South Asian and Sub Saharan countries have values below-average, which is indicated by the negative sign of the respective coefficient estimates. For this fact we can find ample evidence in the literature. Individuals in high-income countries may have a higher ability to pay for personal computers or broadband services (Czernich et al. (2011)) and tend to have higher degree of internet penetration (Hargittai (1999)). The geographical dummy variables approximately match with a high income (Europe) or low (South Asian and Sub Saharan) income levels.

Next, the gross fixed capital formation (gfcf_m_log) also belongs to the group of highly significant variables in both the OLS and KS regression results. The interpretation of the estimation coefficient can be ambiguous. On the one hand, investments in ICT infrastructure are part of the gross fixed capital. Hence, the amount of gross fixed capital formation is increased through higher investments in ICT. On the other hand, investments in certain goods or services increase investments in IT simultaneously. This is in particular the case with goods/services which need ICT infrastructure as a complementary product. These goods/services can be found in smart devices, household electronics, digital media, the automobile industry as well as in industrial products of the mechanical engineering sector or logistics (OECD 2011). Public investment such as the establishment and development of tolling systems or e-government services also require ICT infrastructure as a crucial basis. Despite these obvious relationships, the role of gross fixed capital formation in relation to ICT infrastructure has not been examined widely in the literature.

Also significant in both the OLS and KS regression results is the investment freedom (inv_freedom_m). This variable (provided by the Heritage Foundation) is represented by an index that indicates whether a country allows individuals and firms moving capital across countries' borders, without restriction as well as capital flows internally (score of 100) or with restrictions on investment (score below 100).[27] Countries with a higher score of investment freedom are suggested to attract investors and therefore more (both domestic and foreign direct) investment.[28] As previously mentioned, part of these investments concern products using IT/ICT infrastructure as complementary products.

Only significant in the KS regression is the GDP per person (RGDPP_m_log). As already mentioned in the section above, per capita income was found as the major and mostly identified determinant of ICT in the literature. The fact that the Lasso also selects per capita income to explain global differences in the diffusion of ICT, is assuring for this result.

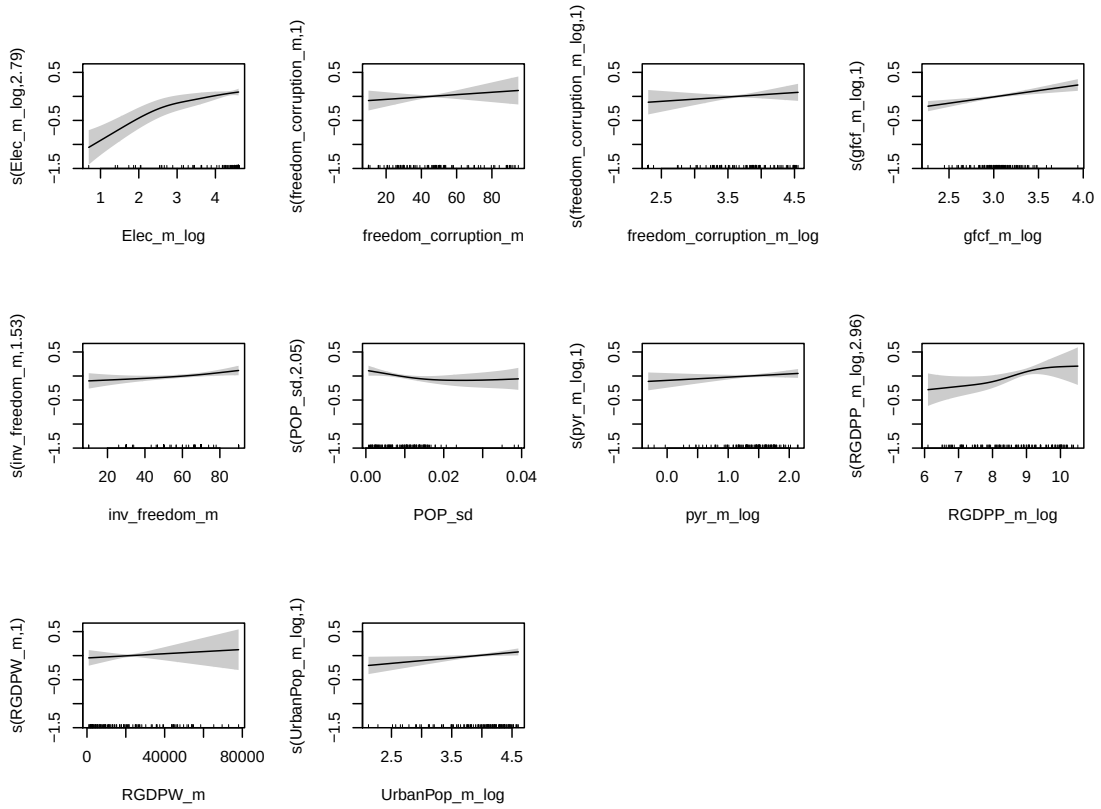
Also only significant in the KS regression is the urban population, measured as percent of total population (UrbanPop_m_log). This result supports the hypothesis, that the urban population tends to adopt more ICT,

[26] Access to electricity is not counted as it does not fit to one of the categories.

[27] Possible restrictions might be rules for foreign and domestic investment, payments, transfers, and capital transactions, restricted access to foreign exchange, labor regulations, corruption, red tape, weak infrastructure, and political and security conditions. For more information on the calculation see the Website of the Heritage Foundation: http://www.heritage.org/index/investment-freedom.

[28] See, for instance, Azman-Saini et al. (2010) for an overview of the discussion, a brief review of the literature and an empirical investigation of the international evidence.

Figure 3: GAM Results for the Three-Stage Procedure (dependent variable is $\ln IT$)



not at least because of possible network economies. The positive effect of urban population is in accordance with the work of Crenshaw and Robinson (2006) as well as Dasgupta et al. (2001).

Also selected but not significant are the dummy variables of Europe and Central Asia, the score of freedom from corruption (freedom_corruption_m(_log)), the standard deviation of the population (POP_sd), the variable of average years of primary schooling attained (pyr_m_log) and the output-side real GDP at chained PPPs per worker (RGDPW_m). Although these variables are not significant in the regressions, the coefficient signs are plausible in this context.

It is remarkable that the Lasso did not select even one human capital variable to explain ICT infrastructure in these regressions. Although we included several variables in the database, none was regarded as a major explanatory variable. This reflects the findings in the literature where human capital is regarded as one of the most relevant ICT explaining factor in theory, whereas empirical evidence could not be established in several studies using various variables.

From the total of 14 selected variables, eight were tested to be significant. From these, three variables describing geographical factors, two describe the economic status and structure and one variable each is included in the categories of demographic factors and regulation. Thus, we find the main areas of relevant influence factors here also well represented. With these variables we can explain about 95 percent of the variation (measured by the adjusted R^2)[29] in the log IT variable, averaged over the period of 2002-2012.

It may be suspected that multicollinearity is a major problem with such a large number of explanatory variables. This is, however, not the case since we find a condition number of about 15 based on the standardized matrix of explanatory variables and there are only very few variance inflation factors which may be viewed as large. This again shows the ability of Lasso-type procedures to successfully avoid multicollinearity.

Associated with the GAM regression results is figure 3 showing the plots of the (centered) spline terms for the selected variables. In the panels of the plot the tick marks at the abscissa (so-called rugs) indicate the positions of the data points of the respective explanatory variable. The gray shaded areas indicate the 95 percent confidence intervals. The equivalent degrees of freedom (edf) values substantially larger than one reveal nonlinear effects of Elec_m_log (access to electricity as percent of population), POP_sd (the standard deviation of population)

[29] In the case of the KS regression, Renaud and Victoria-Feser (2010) explain the kind of R^2 measures used for the assessment of fit.

and RGDPP_m_log (log GDP per person). The other variables appear to have a linear association with $\ln IT$. This assertion can be quickly verified by simply trying to draw a straight line through the gray-shaded 95 percent confidence intervals which actually is possible for the variables deemed linear.

At first in figure 3, the curve of variable Elec_m_log is concavely curved. The nonlinear effect shows that countries with a better electricity supply tend to have a more developed ICT infrastructure but this effect is driven by the large heterogeneity of the electricity supply variable across countries (see the rugs at the bottom of the right-hand panel of the figure). The association is weaker (the curve flatter) for the more advanced countries with a better electricity support system clustered at the upper end of the scale with values above four (approximately corresponds to a 54% access of population to electricity) for this variable. A nonlinear effect of Elec_m_log is not surprising in this case. In general, urban areas were the first to be electrified because many customers hosted on a relatively small area.[30] The share of urban areas to national GDP ranges from an average of 55% in the developing world to 85% in developed countries (Crenshaw and Robinson 2006). In addition, urban population tend to adopt more ICT (internet and computer) because of network economies. For these reasons it can be assumed, that an initial electrification of urban (and mostly more industrialized areas) has a greater impact on the diffusion of ICT infrastructure than an electrification of rural (mostly not industrialized) regions. On closer inspection, the curve of variable Elec_m_log may be decomposed in two straight lines. At the value of approximately 2.5 the curve describes a kink which corresponds to 12% of population having access to electricity. Below this threshold, an increase of electrification has a stronger impact on the level of ICT infrastructure than above.

In the plot of the variable POP_sd, the solid line describes a mildly regressive curve. The rugs show that most of the data points have a value below 0.018 corresponding to a standard deviation about 1. In both theory and literature no connection between ICT infrastructure and the standard deviation of population is discussed. Moreover, the edf value is only slightly larger than 2 we will not further elaborate on this issue.

At last, the edf value of RGDPP_m_log indicates a nonlinear effect. In the plot, the solid curve is s-shaped. As pointed out above, per capita income is an important determinant of computer ownership and internet use (OECD 2001). Hargittai (1999) as well as Beilock and Dimitrova (2003) argue that countries whose citizens are better off economically tend to have more ICT. Based on the assumption that countries with higher per capita income invest more in R&D and therefore are better able to discover and use ICT (Baliamoune-Lutz 2003), per capita income influences the ICT indirect. The curve of variable RGDPP_m_log shows a progressive course up to a level of approximately 9. This value corresponds to an expenditure-side real GDP per person of about 8100 US$. Up to this value, an increase of the GDP per person leads to larger effect on ICT infrastructure. Beyond this level saturation seems to take force.

Since we have dummy variables and other discretely-coded variables in our set of explanatory variables the uniqueness problem raised by Tibshirani (2013) may be an issue. We combat this problem by going a step further and employing the bootstrap Lasso procedure as described above to peel out those explanatory variables which are selected in 90 percent out of 10000 bootstrap replications of the Lasso. This device also delivers us the more robust explanatory variables. As to be expected, we obtain a substantially reduced set of selected variables. The final regression results are reported in table 3 and figure 4.

A first view at the results shows that the explanatory power of these regressions is somewhat reduced but remains well above 0.9. All dummy variables are discarded now by the model selection procedure. The remaining selected variables are all highly significant with one exception (UrbanPop_m_log in the case of the OLS regression). The finding, that GDP per person (RGDPP_m_log) belongs to the group of robust explanatory variables again supports previous results finding per capita income to be a major determinant of ICT.

In contrast to the regression of the three-stage procedure in table 2 the index of freedom from corruption (freedom_corruption_m_log) belongs to the group of significant and even robust explanatory variables explaining global differences in the diffusion of ICT infrastructure. Basic idea of the score is, that "corruption erodes economic freedom by introducing insecurity and uncertainty into economic relationships".[31] The index is provided by the Heritage Foundation and is mainly derived from Transparency International's Corruption Perceptions Index (CPI).[32] Multiplying the CPI by 10, the score of freedom from corruption ranges from 0 (very corrupt government) to 100 (very little corruption). The Heritage Foundation uses qualitative information from internationally recognized and reliable sources to determine freedom from corruption score for countries that are not covered in the CPI.[33] The basic idea of the link between the score of freedom from corruption and ICT

[30] Due to larger distances between customers in few inhabited, rural areas the further electrification causes marginal returns to diminish and so drives average returns down.

[31] Source of the cite: http://www.heritage.org/index/freedom-from-corruption.

[32] The index in turn is composed by several data from various sources. The methodology of the CPI is described by Lambsdorff (2005).

[33] For this purpose they use following sources in order of priority: Transparency International, Corruption Perceptions Index, U.S. Department of Commerce, Country Commercial Guide, Economist Intelligence Unit, Country Commerce, Office of the

Table 3: Regression Results for the bolasso Procedure (dependent variable is $\ln IT$)

	OLS	KS		GAM
c	-0.637	-0.691	c	4.218
	(0.005)	(0.000)		(0.000)
Elec_m_log	0.294	0.275	s(Elec_m_log)	4.352
	(0.000)	(0.000)		(0.000)
freedom_corruption_m_log	0.195	0.174	s(freedom_corruption_m_log)	1.000
	(0.001)	(0.000)		(0.000)
RGDPP_m_log	0.267	0.293	s(RGDPP_m_log)	1.000
	(0.000)	(0.000)		(0.000)
UrbanPop_m_log	0.165	0.161	s(UrbanPop_m_log)	5.922
	(0.117)	(0.009)		(0.002)
R^2	0.916	0.918	R^2	0.939
n	113	113	n	113

Note: Reported are the regression coefficients or the equivalent degrees of freedom (edf) in the case of the spline variables indicated by $s(\cdot)$. Stated in parentheses are p-values of the t-statistics of the F-statistics for the significance of the respective splines.

infrastructure is actually the same as for investment freedom. The fact that corruption in conjunction with the consequences of insecure and uncertain economic relationships, discourages and sometimes prevents investment. Hence, complimentary investments in ICT infrastructure are also not undertaken.

The GAM regression results show that the effects of freedom_corruption_m_log and RGDPP_m_log are clearly linear. As before, nonlinear effects can be uncovered for the variables Elec_m_log and UrbanPop_m_log. For Elec_m_log the associated figure 4 shows a similar curve shape for the bolasso procedure, as for the three-stage procedure. However, at a value about 4 the curve again becomes steeper. In this range of variable values above 4, we have a strong accumulation of rugs. The interpretation of this finding proves difficult, however, because a degressive curve shape (as in figure 3) is more plausible from a theoretical point of view. The reason why the increase of access to electricity above a level of approximately 55% should lead to a larger effect on ICT infrastructure than a level below is not clear and speculative.

The curve of UrbanPop_m_log is shaped like a wave. Different parts of the curve (intervals of 2.6-3.2 and 3.5-4) show that an increase of urban population in these intervals have a greater impact on ICT infrastructure than in the other intervals. It is remarkable that the curve weakly decreases from a value of 4.0 onwards. A further increase of urban population at a level of approximately 55% has a slightly diminishing effect on the level of ICT infrastructure. In this range the number of observations heap up and the gray-shaded 95 percent confidence interval narrows. The diminishing effect of urban population on the level of ICT infrastructure can be explained by congestion effects.

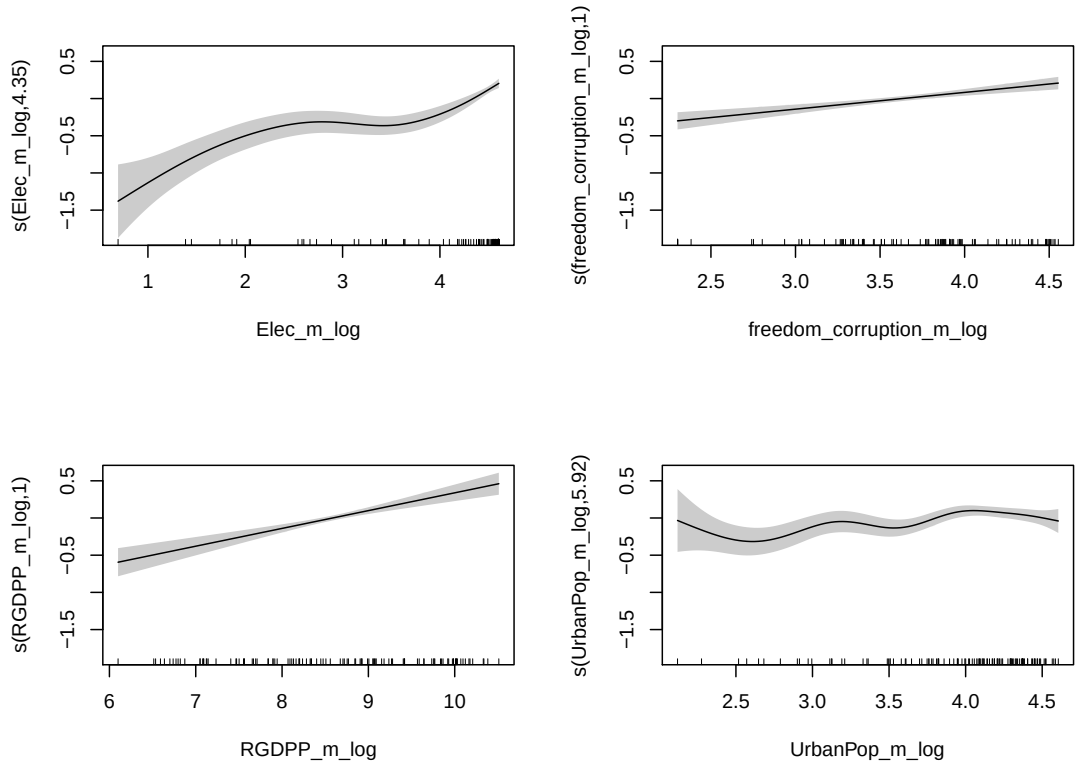
In this subsection we found several variables to explain ICT infrastructure during 2002-2012. In the initial three-stage procedure, a total of 14 variables have been selected explaining about 95 percent of the variation in the log IT variable. Among these variables 8 were tested to be significant with three variables for geographical factors, two describing economic status and structure and one variable are included in each of the categories of demographic factors, regulation and miscellaneous. In the associated GAM regression we could reveal nonlinear effects of Elec_m_log, POP_sd and RGDPP_m_log. In the subsequent bootstrap Lasso (bolasso) procedure we get a reduced set of more robust explanatory variables. This procedure selects four variables with an explanatory power of well above 0.9 in all regressions. Furthermore assuring is the close correspondence of the OLS and the robust KS regression estimates. The following GAM regression detects nonlinear effects of Elec_m_log and UrbanPop_m_log.

Our results show the power of a wide variety of variables for explaining cross-country differences in ICT infrastructure. We are able to reestablish the empirical evidence for per capita income as significant and robust variable describing global differences in ICT infrastructure. Although we examined many empirical studies, we could not find evidence for Elec_m_log in the literature. This issue is quite remarkable as electricity can be viewed as a fundamental infrastructural prerequisite for ICT. This is even surprising as we found Elec_m_log to be a very robust and significant explanatory variable across all estimates.

5.2 Explaining ICT Infrastructure During 2008-2012

U.S. Trade Representative, National Trade Estimate Report on Foreign Trade Barriers; and official government publications of each country.

Figure 4: GAM Results for the bolasso Procedure (dependent variable is $\ln IT$)



Turning to the results with $\ln IT_2$ as the dependent variable we find that only two explanatory variables are selected by the three-stage procedure. Recall that $\ln IT_2$ is the log average over the period 2008-2012. This allows is to put $\ln IT_1$ (the log average over the previous period 2002-2006) into the set of candidate explanatory variables. As shown in table 4 we see that the ICT infrastructure variable is characterized by persistence since $\ln IT_1$ appears as a strongly significant explanatory variable associated with a positive coefficient estimate. The coefficient estimate of about 0.34 (smaller than one) is indicative for the presence of conditional convergence of the ICT infrastructure across countries.[34] It seems that this persistence captures almost the entire amount of explanatory power of the other variables which were selected in the previous subsection. An exception is the variable Elec_m_log which remains strongly significant with a positive coefficient estimate although of a reduced magnitude. The coefficient estimates with the OLS and the robust KS estimates are rather similar, the overall explanatory power is also remains substantial.

The GAM estimates point to a linear influence of log_IT1 and a nonlinear effect of Elec_m_log as can be seen from figure 5. The nonlinear effect shows a similar degressive course as in figure 2, again indicating that countries with a better electricity supply tend to have a more developed ICT infrastructure.

The results of the bolasso procedure lead to exactly the same variable selection and therefore to the same results as the three-stage procedure. Therefore, we need not show the corresponding table and figure at this point. While 14 variables (among 9 significant) were necessary to obtain an explanatory power of about 0.9 for the regressions of period 2002-2012, only two variables already achieve an explanatory power of 0.92 for the subperiod of 2008-2012. Clearly, much of the effects already incorporated in the lagged IT variable as an explanatory variable.

6 Conclusion

The analysis discussed in this paper reveals that a set of explanatory variables, selected from a wide array of candidate variables, is very well able to explain cross-country differences in ICT infrastructure for a broad

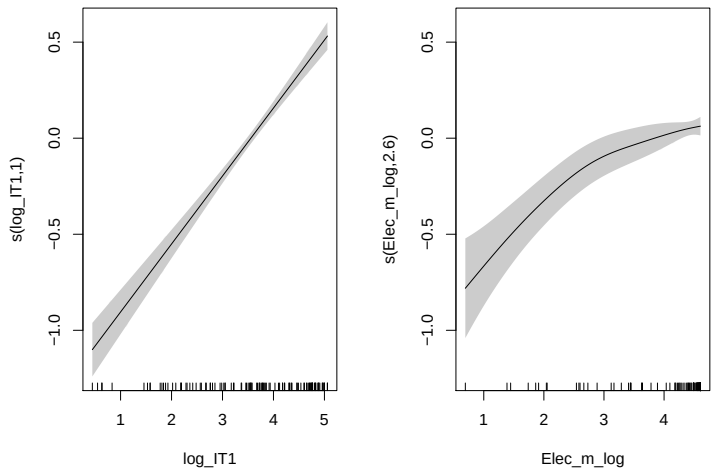
[34] Subtracting $\ln IT_1$ from both sides results in the change of $\ln IT$ on the left-hand side and a negative coefficient $0.34 - 1$ on the right-hand side. This is the indication of conditional convergence investigated in cross-section growth empirics, see e.g. Barro and Sala-i-Martin (1991, 1992).

Table 4: Regression Results for the Three-Stage Procedure (dependent variable is $\ln IT_2$)

	OLS	KS		GAM
c	2.641	2.695	c	4.557
	(0.000)	(0.000)		(0.000)
log_IT1	0.337	0.344	s(log_IT1)	1.000
	(0.000)	(0.000)		(0.000)
Elec_m_log	0.175	0.155	s(Elec_m_log)	2.601
	(0.000)	(0.000)		(0.000)
R^2	0.920	0.919	R^2	0.926
n	113	113	n	113

Note: Reported are the regression coefficients or the equivalent degrees of freedom (edf) in the case of the spline variables indicated by $s(\cdot)$. Stated in parentheses are p-values of the t-statistics of the F-statistics for the significance of the respective splines.

Figure 5: GAM Results for the Three-Stage Procedure (dependent variable is $\ln IT_2$)



sample of more than 100 countries of all stages of development. The dependent variable for ICT infrastructure is constructed from a principal components analysis. As discussed, this variable could alternatively be interpreted as a proxy for either ICT equipment or ICT usage because of the close relation of both aspects.

We can in particular show that real income per capita, electricity usage, urbanization, indicators of regulatory and institutional aspects as well as regional dummies are major determinants of ICT infrastructure. The explanatory variables are selected from a broad set of candidate variables by variants of the Lasso approach which have been developed in the machine learning literature. Jointly, these variables achieve a very high degree of explanatory power. We find the results to be robust to heteroskedasticity and outlying observations. The former is assured by using a heteroskedasticity correction of the standard errors, while the latter is checked by comparing the least squares coefficient estimates to those of a robust regression estimator. We also applied a semiparametric GAM estimator and uncovered nonlinear effects for some explanatory variables, i.e. electricity usage. The bulk of the explanatory power, however, stems from the linear effects of the regressors.

The findings regarding the electricity variable are particularly interesting from the perspective of growth economics. This variable is very robustly selected by the different approaches pursued. Comin and Hobijn (2004) highlight electricity production as an important prerequisite for the adoption of other technologies. Electricity is therefore a genuine general purpose technology in the sense Bresnahan and Trajtenberg (1995), characterized by its pervasiveness and its role as a central precondition for other technologies. One sector which is particularly depending on electricity is the entire ICT sector.

In a further analysis, splitting the sample period into two subperiods we can also establish conditional convergence of the ICT variable. This may be taken as evidence against the global digital divide. Interestingly, human capital indicators are not selected, although many of them are included in the set of candidate variables. Thus, human capital differences across countries seem not to be directly related to differences in ICT infrastructure. At a first view this seems counter-intuitive but it may be explained by the fact that many end devices are so easy to operate that not much formal education is actually needed for their usage. For the setup of the infrastructure only a few specialists are required which may also be hired from abroad.

This research can be improved along several lines which are avenues for future research. It is clearly worth the effort to improve the construction of the dependent variable for ICT infrastructure. Since we have exploited the publicly available data sources to a considerable extent with an eye on reaching a broad cross-country sample, this would require making use of information from commercial sources which is available, e.g., from the International Telecommunication Union (ITU). Moreover, basically the data could be treated as a panel and panel regression estimators could be employed. We abstained from picking-up this possibility in the present paper since the period for which the indicators required for the construction of the dependent variable is rather short. Performing an analysis aiming to assess the long-run determinants of ICT infrastructure could be easily obscured by using a panel on a year-by-year basis emphasizing the role of short-run fluctuations. Finally, the country coverage or the set of available variables could be increased by trying to apply imputation methods for closing gaps in the available data. The sample of available countries is nevertheless reasonably large even without imputation which would also generate additional noise in the analysis.

Appendix

Table 5: Correlation Coefficients of ICT Variables

	(1)	(2)	(3)	(4)	(5)
(1) Telephone Lines	1.00				
(2) Internet Users	0.91	1.00			
(3) Broadband Internet Subscribers	0.90	0.92	1.00		
(4) Mobile Cell Subscribers	0.77	0.80	0.70	1.00	
(5) PCs	0.83	0.87	0.87	0.65	1.00

Note. All Variables are in values per 100 people and averaged over the years of 2002-2012. Pearson correlation coefficients are computed between each pair of variables using all complete pairs of observations on those variables. Based on 204 observations, the correlation coefficients of 179 complete pairs are computed.

Figure 6: Density Plots of Dependent Variables

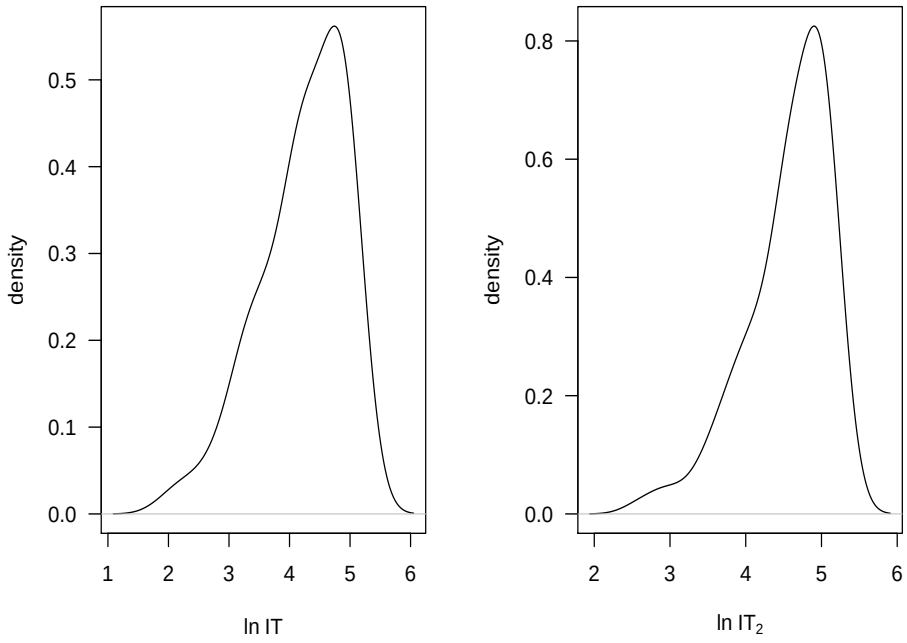


Table 6: Definition and Sources of Explanatory Variables

Variable	Source	Available at	Category	Literature references	Transform.		
					M	G	SD
Dummy variable for advanced countries	Barro-Lee	Barro-Lee	Economic Status and Structure				
Agriculture share in GDP	World Bank national accounts data, and OECD National Accounts data files.	World Bank	..	Caselli and Coleman (2001)	x	x	x
Share of gross capital formation (at current PPPs)	PWT	PWT	..		x	x	x
Developing countries	IMF	IMF	..				
Developed countries	IMF	IMF	..				
Number of persons engaged (in millions)	PWT	PWT	..		x	x	x
Foreign direct investments	International Monetary Fund, International Financial Statistics and Balance of Payments databases, World Bank, International Debt Statistics, and World Bank and OECD GDP estimates	World Bank	..	Crenshaw and Robinson (2006)	x	x	
Private credit by deposit money banks to GDP (%)	International Financial Statistics (IFS), International Monetary Fund (IMF)	World Bank	..		x	x	
Gross fixed capital formation (% of GDP and constant 2005 US$)	World Bank national accounts data, and OECD National Accounts data files	World Bank	..		x	x	
Gini coefficient (World Bank estimate)	World Bank, Development Research Group	World Bank	..	Wunnava and Leiter (2009)	x		
Gross national expenditure (% of GDP)	World Bank national accounts data, and OECD National Accounts data files	World Bank	..		x		
Capital stock	PWT	own calculation	..		x	x	x
Economic Globalization	Axel Dreher	QOG	..		x	x	
Index of Globalization	Axel Dreher	QOG	..		x	x	
Commercial banks and other lending (PPG + PNG) (NFL, current US$)	World Bank, International Debt Statistics	World Bank	..		x		
Manufacturing, value added (% of GDP)	World Bank national accounts data, and OECD National Accounts data files	World Bank	..		x	x	
Market Capitalization of listed Companies (% of GDP)	Standard & Poor's, Global Stock Markets Factbook and supplemental S&P data	World Bank	..		x		

Variable	Source	Available at	Category	Literature references	Transform.		
					M	G	SD
Colony Dummy	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Expenditure-side real GDP at chained PPPs (in mil. 2005US$) per person	PWT	PWT	..	most of found literature	x	x	x
Output-side real GDP at chained PPPs (in mil. 2005US$) per worker	PWT	PWT	..	most of found literature	x	x	x
employment of service sector (% of total)	International Labour Organization, Key Indicators of the Labour Market database	World Bank	..	Gust and Marquez (2004)	x		
Services, etc., value added (% of GDP)	World Bank national accounts data, and OECD National Accounts data files	World Bank	..		x	x	
Adult literacy rate	UNESCO Institute for Statistics	World Bank	Human Capital	Baliamoune-Lutz (2003)	x		
Percentage of graduates from tertiary education graduating from Engineering, Manufacturing and Construction programmes	UNESCO Institute for Statistics	World Bank	..		x		
Index of human capital per person	PWT	PWT	..		x	x	
Average Years of Tertirary Schooling Attained	Barro-Lee	Barro-Lee	..		x	x	
Average Years of Primary Schooling Attained	Barro-Lee	Barro-Lee	..		x	x	
Researchers in R&D (per million people)	United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics	World Bank	..		x		
Research and development expenditure (% of GDP)	United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute for Statistics	World Bank	..		x		
Percentage of graduates from tertiary education graduating from Science programmes	UNESCO Institute for Statistics	World Bank	..		x		
Percentage of graduates from tertiary education graduating from Services programmes	UNESCO Institute for Statistics	World Bank	..		x		
Average Years of Secondary Schooling Attained	Barro-Lee	Barro-Lee	..		x	x	

Variable	Source	Available at	Category	Literature references	Transform.		
					M	G	SD
Colony Dummy	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
tertiary education enrollment	UNESCO Institute for Statistics	World Bank	..	Crenshaw and Robinson (2006)	x	x	
Average Years of Schooling Attained	Barro-Lee	Barro-Lee	..		x	x	
English Speaking Population	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Fraction Speaking Foreign Language	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Business Freedom	Heritage Foundation	Heritage Foundation	Regulation		x		
Control of Corruption	World Bank (Worldwise Gov. Ind.)	World Bank	..		x		
CPIA business regulatory environment rating (1=low to 6=high)	World Bank Group, CPIA database	World Bank	..		x		
CPIA social protection rating (1=low to 6=high)	World Bank Group, CPIA database	World Bank	..		x		
Index of Economic Freedom	Heritage Foundation	Heritage Foundation	..	Baliamoune-Lutz (2003)	x		
Political Rights	Freedom House	QOG	..	Baliamoune-Lutz (2003), Wunnava and Leiter (2009)	x		
Financial Freedom	Heritage Foundation	Heritage Foundation	..		x		
Fiscal Freedom	Heritage Foundation	Heritage Foundation	..		x		
Freedom from Corruption	Heritage Foundation	Heritage Foundation	..		x		
Government Effectiveness	World Bank (Worldwise Gov. Ind.)	World Bank	..		x		
ICRG Indicator of Quality of Government	International Country Risk Guide/The PRS Group	QOG	..		x		
Investment Freedom	Heritage Foundation	Heritage Foundation	..		x		
Monetary Freedom	Heritage Foundation	Heritage Foundation	..		x		
Property Rights	Heritage Foundation	Heritage Foundation	..	Crenshaw and Robinson (2006), Caselli and Coleman (2001)	x		
Political Stability and Absence of Violence/Terrorism	World Bank (Worldwise Gov. Ind.)	World Bank	..		x		

Variable	Source	Available at	Category	Literature references	Transform.		
					M	G	SD
Colony Dummy	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Rule of Law	World Bank (Worldwise Gov. Ind.)	World Bank	..		x		
Regulatory Quality	World Bank (Worldwise Gov. Ind.)	World Bank	..	Chinn and Fairlie (2007)	x		
Corruption Perceptions Index	Transparency International	QOG	..		x		
Trade Freedom	Heritage Foundation	Heritage Foundation	..		x		
Voice and Accountability	World Bank (Worldwise Gov. Ind.)	World Bank	..		x		
Control of Corruption - Estimate	The Worldwide Governance Indicators	QOG	..		x		
Political Stability - Estimate	The Worldwide Governance Indicators	QOG	..		x		
Rule of Law - Estimate	The Worldwide Governance Indicators	QOG	..		x		
Civil Liberties	Freedom House	QOG	Demographic Factors	Baliamoune-Lutz (2003), Beilock and Dimitrova 2003), Wunnava and Leiter (2009)	x		
Population (in millions)	PWT	PWT	..		x	x	x
Dummy variable for East Asian and Pacific countries	Barro-Lee	Barro-Lee	Geographical Factors	Dasgupta et al. (2001)			
Countries of the European Union (Dummy)			..				
Dummy variable for countries in Europe and Central Asia	Barro-Lee	Barro-Lee	..				
Dummy variable for Latin American and Carribean countries	Barro-Lee	Barro-Lee	..	Dasgupta et al. (2001)			
	Barro-Lee	Barro-Lee	..				
Countries of the OECD (Dummy)			..				
Population of largest city as % of world urban total	United Nations, World Urbanization Prospects.	World Bank	..	Crenshaw and Robinson (2006)	x	x	
Dummy for South Asian countries	Barro-Lee	Barro-Lee	..				
Dummy for Sub-Saharan Africa	Barro-Lee	Barro-Lee	..	Dasgupta et al. (2001)			
Urban population (% of total)	United Nations, World Urbanization Prospects.	World Bank	..	Dasgupta et al. (2001)	x	x	
Air Distance to Big Cities	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				

Variable	Source	Available at	Category	Literature references	Transform.		
					M	G	SD
Colony Dummy	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Catholics as percentage of population in 1980	La Porta, Lóp ez-de-Silanes, Shleifer and Vishny	QOG	Miscallaneous		x		
Access to electricity (% of population)	World Bank, Sustainable Energy for all (SE4ALL) database from World Bank, Global Electrification database	World Bank	..		x		
Muslims as percentage of population in 1980	La Porta, Lóp ez-de-Silanes, Shleifer and Vishny	QOG	..				
Protestants as percentage of population in 1980	La Porta, Lóp ez-de-Silanes, Shleifer and Vishny	QOG	..		x		
British Colony Dummy	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Socialist Dummy	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Spanish Colony	Sala-i-Martin/Doppelhofer/Miller	SiMDM	..				
Dummy variable for former Spanish colonies	Barro (1999)	SiMDM	..				

Note: The table contains all data collected in the database. The abbreviation SiMDM stands for Sala-i-Martin, Doppelhofer and Miller. In column 5 we describe whether and where the variables are mentioned in the literature. The last three columns indelicate, whether the mean value (M), growth rate (G) and/or standard derivation of the growth rate (sd) has been calculated for a specific variable.

Table 7: List of Countries

Country code	IT data complete	in final dataset	Country code	IT data complete	in final dataset	Country code	IT data complete	in final dataset
ABW			GIN	x		NIC	x	
AFG	x		GMB	x		NLD	x	x
AGO	x		GNB	x		NOC		
ALB	x	x	GNQ	x		NOR	x	x
AND			GRC	x	x	NPL	x	x
ARB			GRD	x		NZL	x	x
ARE	x		GRL			OEC		
ARG	x	x	GTM	x	x	OED		
ARM	x	x	GUM			OMN	x	
ASM			GUY	x		OSS		
ATG	x		HIC			PAK	x	x
AUS	x	x	HKG	x		PAN	x	x
AUT	x	x	HND	x	x	PER	x	x
AZE	x		HPC			PHL	x	x
BDI	x	x	HRV	x	x	PLW		
BEL	x	x	HTI	x		PNG	x	
BEN	x	x	HUN	x	x	POL	x	x
BFA			IDN	x	x	PRI		
BGD	x	x	IMN			PRK		
BGR	x	x	IND	x	x	PRT	x	x
BHR	x	x	INX			PRY	x	x
BHS	x		IRL	x	x	PSE		
BIH	x		IRN	x	x	PSS		
BLR			IRQ	x		PYF		
BLZ	x		ISL	x	x	QAT	x	x
BMU			ISR	x	x	REU		
BOL	x	x	ITA	x	x	ROU	x	
BRA	x	x	JAM	x	x	RUS	x	x
BRB	x		JOR	x	x	RWA	x	x
BRN	x		JPN	x	x	SAS		
BTN	x		KAZ			SAU	x	x
BWA	x	x	KEN	x	x	SDN	x	
CAA			KGZ	x	x	SEN	x	x
CAF	x		KHM	x	x	SER		
CAN	x	x	KIR	x		SGP	x	x

Country code	IT data complete	in final dataset	Country code	IT data complete	in final dataset	Country code	IT data complete	in final dataset
CEA			KNA	x		SLB	x	
CEU			KOR	x	x	SLE		
CHE	x	x	KSV			SLV	x	x
CHI			KWT	x	x	SMR	x	
CHL	x	x	LAC			SOM	x	
CHN	x	x	LAO	x	x	SRB	x	
CIV	x	x	LBN	x		SSA		
CLA			LBR			SSD		
CME			LBY	x		SSF		
CMR	x	x	LCA	x		SST		
COD			LCN			STP	x	
COG	x	x	LDC			SUR	x	
COL	x	x	LIC			SVK	x	x
COM	x		LIE			SVN	x	x
CPV	x		LKA	x	x	SWE	x	x
CRI	x	x	LMC			SWZ	x	x
CSA			LMY			SXM		
CSS			LSO	x	x	SYC	x	
CUB	x		LTU	x	x	SYR	x	x
CUW			LUX	x	x	TCA		
CYM			LVA	x	x	TCD	x	
CYP	x	x	MAC			TGO	x	x
CZE	x	x	MAF			THA	x	x
DEU	x	x	MAR	x	x	TJK	x	
DJI	x		MCO			TKM	x	
DMA	x		MDA	x	x	TLS		
DNK	x	x	MDG	x		TON	x	
DOM	x	x	MDV	x		TTO	x	x
DZA	x		MEA			TUN	x	x
EAP			MEX	x	x	TUR	x	x
EAS			MHL	x		TUV	x	
ECA			MIC			TWN		
ECS			MKD	x		TZA	x	x
ECU	x	x	MLI	x	x	UGA	x	x
EGY	x	x	MLT			UKR	x	x
EMU			MMR	x		UMC		
ERI	x		MNA			URY	x	x
ESP	x	x	MNE			USA	x	x

Country code	IT data complete	in final dataset	Country code	IT data complete	in final dataset	Country code	IT data complete	in final dataset
EST	x	x	MNG	x	x	UZB	x	
ETH	x		MNP			VCT	x	
EUU			MOZ	x	x	VEN	x	x
FIN	x	x	MRT	x	x	VIR		
FJI	x		MUS	x	x	VNM	x	x
FRA	x	x	MWI	x	x	VUT	x	
FRO			MYS	x	x	WLD		
FSM	x		NAC			WSM	x	
GAB	x	x	NAM	x	x	YEM	x	
GBR	x	x	NCL			ZAF	x	x
GEO	x		NER	x	x	ZMB	x	x
GHA	x	x	NGA	x		ZWE	x	x

Note: The column "Country code" lists all available countries by ISO ALPHA-3 code, for which data is available in the database. In column "IT complete" all countries with data for IT infrastructure are listed. The column "in final dataset" lists all countries, included in the final dataset.

Table 8: Descriptive Statistics

Variable	Description	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Advanced Economies	Dummy variable for advanced countries	0.000	0.000	0.000	0.212	0.000	1.000
Business Freedom	Business Freedom	40.000	55.000	70.000	67.030	73.000	100.000
Business Freedom (log)	Business Freedom	3.689	4.007	4.248	4.186	4.290	4.605
catholic_m	Catholics as percentage of population in 1980	0.000	0.800	18.700	34.600	76.400	96.900
CSH_m	Share of gross capital formation (at current PPPs)	0.051	0.139	0.194	0.199	0.249	0.522
CSH_m_log	Share of gross capital formation (at current PPPs)	-2.968	-1.974	-1.641	-1.700	-1.392	-0.650
CSH_sd	Share of gross capital formation (at current PPPs)	0.030	0.072	0.110	0.140	0.181	0.493
East Asia and the Pacific	Dummy variable for East Asian and Pacific countries	0.000	0.000	0.000	0.097	0.000	1.000
econ_freedom_m	Index of Economic Freedom	35.900	54.100	61.080	60.380	67.280	86.950
econ_freedom_m_log	Index of Economic Freedom	3.581	3.991	4.112	4.086	4.209	4.465
Elec_m	Access to electricity (% of population)	2.000	65.500	93.820	75.650	100.000	100.000
Elec_m_log	Access to electricity (% of population)	0.693	4.182	4.541	4.104	4.605	4.605
EMP_m	Number of persons engaged (in millions)	0.134	1.496	3.418	17.600	10.170	605.900
EMP_m_log	Number of persons engaged (in millions)	-2.007	0.403	1.229	1.325	2.320	6.407
EMP_sd	Number of persons engaged (in millions)	0.005	0.015	0.019	0.021	0.023	0.066
EU	Countries of the European Union (Dummy)	0.000	0.000	0.000	0.195	0.000	1.000
Europe and Central Asia	Dummy variable for countries in Europe and Central Asia	0.000	0.000	0.000	0.142	0.000	1.000
fh_cl_m	Civil Liberties	1.000	2.190	3.810	3.608	4.810	6.952
fh_cl_m_log	Civil Liberties	0.000	0.784	1.338	1.144	1.571	1.939
fh_pr_m	Political Rights	1.000	1.333	3.429	3.400	4.905	6.905
fh_pr_m_log	Political Rights	0.000	0.288	1.232	1.026	1.590	1.932
financ_freedom_m	Financial Freedom	10.000	50.000	50.000	54.220	66.670	90.000
financ_freedom_m_log	Financial Freedom	2.303	3.912	3.912	3.916	4.200	4.500

Variable	Description	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fiscal_freedom_m	Fiscal Freedom	30.700	54.600	67.450	65.530	76.980	99.900
freedom_corruption_m	Freedom from Corruption	10.000	27.830	46.000	45.190	57.170	95.000
freedom_corruption_m_log	Freedom from Corruption	2.303	3.326	3.829	3.632	4.046	4.554
gfcf_m	Gross fixed capital formation (% of GDP and constant 2005 US$)	9.629	18.560	20.880	21.600	23.830	51.390
gfcf_m_log	Gross fixed capital formation (% of GDP and constant 2005 US$)	2.265	2.921	3.039	3.040	3.171	3.939
gneGDP_m	Gross national expenditure (% of GDP)	83.060	98.270	101.500	104.300	108.300	212.500
gneGDP_m_log	Gross national expenditure (% of GDP)	4.420	4.588	4.620	4.641	4.685	5.359
HC_m	Index of human capital per person	1.126	1.836	2.212	2.242	2.686	3.429
HC_m_log	Index of human capital per person	0.118	0.607	0.794	0.774	0.988	1.232
hyr_m	Average Years of Tertirary Schooling Attained	0.006	0.082	0.236	0.280	0.400	1.210
hyr_m_log	Average Years of Tertirary Schooling Attained	-5.116	-2.501	-1.444	-1.739	-0.916	0.191
inv_freedom_m	Investment Freedom	10.000	50.000	63.330	58.190	70.000	90.000
inv_freedom_m_log	Investment Freedom	2.303	3.912	4.148	4.010	4.248	4.500
K_m	Capital stock at chained PPPs (in mil. 2005US$)	5,112.000	30,270.000	93,670.000	886,300.000	529,300.000	24,720,000.000
K_m_log	Capital stock at chained PPPs (in mil. 2005US$)	8.539	10.320	11.450	11.820	13.180	17.020
K_sd	Capital stock at chained PPPs (in mil. 2005US$)	0.014	0.028	0.040	0.048	0.057	0.169
Latin America and the Caribbean	Dummy variable for Latin American and Carribean countries	0.000	0.000	0.000	0.168	0.000	1.000
Middle East and North Africa		0.000	0.000	0.000	0.106	0.000	1.000
monet_freedom_m	Monetary Freedom	17.050	63.180	70.350	68.420	81.030	91.000
muslim_m	Muslims as percentage of population in 1980	0.000	0.000	1.000	18.520	16.200	99.400
OECD	Countries of the OECD (Dummy)	0.000	0.000	0.000	0.301	1.000	1.000
POP_m	Population (in millions)	0.252	3.551	8.680	40.850	24.780	1,101.000
POP_m_log	Population (in millions)	-1.379	1.267	2.161	2.273	3.210	7.004
POP_sd	Population (in millions)	0.001	0.005	0.010	0.010	0.014	0.039

Variable	Description	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
property_rights_m	Property Rights	10.000	50.000	53.330	58.400	70.000	90.000
property_rights_m_log	Property Rights	2.303	3.912	3.977	3.981	4.248	4.500
protestant_m	Property Rights	0.000	0.200	1.900	12.880	16.100	97.800
pyr_g	Average Years of Primary Schooling Attained	-0.356	0.060	0.229	0.254	0.423	1.093
pyr_m	Average Years of Primary Schooling Attained	0.734	3.230	4.304	4.222	5.268	8.542
pyr_m_log	Average Years of Primary Schooling Attained	-0.309	1.172	1.460	1.353	1.662	2.145
RGDPP_m	Expenditure-side real GDP at chained PPPs (in mil. 2005US$) per person	445.900	2,140.000	6,035.000	9,286.000	15,400.000	36,710.000
RGDPP_m_log	Expenditure-side real GDP at chained PPPs (in mil. 2005US$) per person	6.100	7.669	8.705	8.585	9.642	10.510
RGDPP_sd	Expenditure-side real GDP at chained PPPs (in mil. 2005US$) per person	0.016	0.033	0.046	0.053	0.061	0.256
RGDPW_m	Output-side real GDP at chained PPPs (in mil. 2005US$) per worker	913.000	6,337.000	16,730.000	22,140.000	35,910.000	78,080.000
RGDPW_m_log	Output-side real GDP at chained PPPs (in mil. 2005US$) per worker	6.817	8.754	9.725	9.530	10.490	11.270
RGDPW_sd	Output-side real GDP at chained PPPs (in mil. 2005US$) per worker	0.013	0.031	0.048	0.056	0.065	0.283
South Asia	Dummy for South Asian countries	0.000	0.000	0.000	0.044	0.000	1.000
Sub-Saharan Africa	Dummy for Sub-Saharan Africa	0.000	0.000	0.000	0.230	0.000	1.000
syr_g	Average Years of Secondary Schooling Attained	-0.086	0.308	0.562	0.559	0.793	2.144
syr_m	Average Years of Secondary Schooling Attained	0.086	1.100	2.016	2.136	3.068	5.198
syr_m_log	Average Years of Secondary Schooling Attained	-2.453	0.095	0.701	0.519	1.121	1.648
trade_freedom_m	Trade Freedom	14.000	55.400	65.070	62.340	76.000	83.000
tyr_g	Average Years of Schooling Attained	-0.260	0.205	0.321	0.355	0.492	1.099

Variable	Description	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
tyr_m	Average Years of Schooling Attained	0.906	4.758	6.600	6.638	8.600	12.310
tyr_m_log	Average Years of Schooling Attained	-0.099	1.560	1.887	1.783	2.152	2.510
UrbanPop_g	Urban population (% of total)	-0.079	0.023	0.095	0.161	0.242	1.487
UrbanPop_m	Urban population (% of total)	8.321	40.410	60.100	57.160	76.030	100.000
UrbanPop_m_log	Urban population (% of total)	2.119	3.699	4.096	3.924	4.331	4.605
wbgi_corcon_m	Control of Corruption - Estimate	-1.186	-0.615	-0.190	0.186	0.717	2.441
wbgi_pse_m	Political Stability - Estimate	-2.275	-0.583	-0.026	0.040	0.746	1.540
wbgi_rle_m	Rule of Law - Estimate	-1.587	-0.606	-0.066	0.133	0.839	1.930
$\ln IT$	IT infrastructure (average over 2002-12)	1.972	3.797	4.322	4.218	4.828	5.166
$\ln IT_1$	IT infrastructure (average over 2002-06)	0.446	2.757	3.768	3.556	4.657	5.059
$\ln IT_2$	IT infrastructure (average over 2008-12)	2.597	4.265	4.654	4.557	4.975	5.254

Note: The Suffix '_m' denotes that the variables is averaged over the years of 1980 to 2000. Accordingly '_m_log' denotes logarithm of the averaged value, '_g' the growth rate and '_sd' the standard derivation of the specific variable.

References

[1] Azman-Saini, W.N.W., Baharumshah, A.Z., Law, S.H. (2010), Foreign Direct Investment, Economic Freedom and Economic Growth: International Evidence, Economic Modelling 27, 1079-1089.

[2] Bach, F.R. (2008), Bolasso: Model Consistent Lasso Estimation Through the Bootstrap, Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland.

[3] Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M. (2015a), Demand Estimation with Machine Learning and Model Combination, NBER Working Paper 20955.

[4] Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M. (2015b), Machine Learning Methods for Demand Estimation, American Economic Review: Papers and Proceedings 105, 481-485.

[5] Baliamoune-Lutz, M. (2003), An Analysis of the Determinants and Effects of ICT Diffusion in Developing Countries, Information Technology for Development 10, 151-169.

[6] Barro, R.J., Lee, J.-W. (2000), International Data on Educational Attainment Updates and Implications, NBER Working Paper 7911.

[7] Barro, R.J., Lee, J.-W. (2013), A New Data Set of Educational Attainment in the World, 1950-2010, Journal of Development Economics 104, 184-198.

[8] Barro, R.J., Sala-i-Martin, X. (1991), Convergence Across States and Regions, Brookings Papers on Economic Activity, 107-182.

[9] Barro, R.J., Sala-i-Martin, X. (1992), Convergence, Journal of Political Economy 100, 223-251.

[10] Bayo-Moriones, A., Lera-López, F. (2007), A Firm-Level Analysis of Determinants of ICT Adoption in Spain, Technovation 27, 352-366.

[11] Behrman, J.R., Rosenzweig, M.R. (1994), Caveat Emptor: Cross-Country Data on Education and the Labour Force, Journal of Development Economics 44, 147-171.

[12] Beilock, R., Dimitrova, D.V. (2003), An Exploratory Model of Inter-Country Internet Diffusion, Telecommunications Policy 27, 237-252.

[13] Belloni, A., Chen, D., Chernozhukov, V., Hansen, C. (2012), Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain, Econometrica 80, 2369-2429. #

[14] Bresnahan, T.F., Trajtenberg, M. (1995), General Purpose Technologies: Engines of Growth?, Journal of Econometrics 65, 83-108.

[15] Caselli, F., Coleman, W.J. (2001), Cross-Country Technology Diffusion: The Case of Computers, American Economic Review 91, 328-335.

[16] Chernozhukov, V., Hansen, C., Spindler, M. (2015), Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments, American Economic Review: Papers and Proceedings 105, 486-490.

[17] Chinn, M.D., Fairlie, R.W. (2007), The Determinants of the Global Digital Divide: A Cross-Country Analysis of Computer and Internet Penetration, Oxford Economic Papers 59, 16-44.

[18] Chinn, M.D., Fairlie, R.W. (2010), ICT Use in the Developing World: A Review of International Economics 18, 153-167.

[19] Comin, D., Hobijn, B. (2004), Cross-Country Technology Adoption: Making the Theories Face the Facts, Journal of Monetary Economics 51, 39-83.

[20] Crenshaw, E.M., Robison, K.K. (2006), Globalization and the Digital Divide: The Roles of Structural Conduciveness and Global Connection in Internet Diffusion, Social Science Quarterly 87, 190-207.

[21] Czernich, N., Falck, O., Kretschmer, T., and Woessmann, L. (2011), Broadband Infrastructure and Economic Growth, Economic Journal 552, 505-532.

[22] Dasgupta, S., Lall, S., Wheeler, D. (2001), Policy Reform, Economic Growth, and the Digital Divide: An Econometric Analysis, Policy Research Working Paper, Vol. 256, World Bank Publications.

[23] Dasgupta, S., Lall, S., Wheeler, D. (2005), Policy Reform, Economic Growth and the Digital Divide, Oxford Development Studies 33, 229-243.

[24] Doornik, J.A., Hendry, D.F. (2015), Statistical Model Selection with "Big Data", Cogent Economics and Finance 3, 1045216.

[25] Guerrieri, P., Luciani, M., Meliciani, V. (2011), The Determinants of Investment in Information and Communication Technologies, Economics of Innovation and New Technology 20, 387-403.

[26] Guillén, M.F., Suárez, S.L. (2001), Developing the Internet: Entrepreneurship and Public Policy in Ireland, Singapore, Argentina, and Spain, Telecommunications Policy 25, 349-371.

[27] Guillén, M.F., Suárez, S.L. (2005), Explaining the Global Digital Divide: Economic, Political and Sociological Drivers of Cross-National Internet Use, Social Forces 84, 681-708.

[28] Gust, C., Marquez, J. (2004), International Comparisons of Productivity Growth: The Role of Information Technology and Regulatory Practices, Labour Economics 11, 33-58.

[29] Haller, S., Traistaru-Siedschlag, J. (2007), The Adoption of ICT: Firm-Level Evidence from Irish Manufacturing Industries, Papers WP204, Economic and Social Research Institute (ESRI).

[30] Hargittai, E. (1999), Weaving the Western Web: Explaining Differences in Internet Connectivity Among OECD Countries, Telecommunications Policy 23, 701-718.

[31] Hargittai, E. (2003), The Digital Divide and What to Do About It, New Economy Handbook, 821-839.

[32] Hoerl, A.E., Kennard, R.W. (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics 12, 55-67.

[33] Jorgenson, D.W. Stiroh, K.J. (1999), Information Technology and Growth, American Economic Review 89, 109-115.

[34] Kiiski, S., Pohjola, M. (2002), Cross-Country Diffusion of the Internet, Information Economics and Policy 14, 297-310.

[35] Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z. (2015), Prediction Policy Problems, American Economic Review: Papers and Proceedings 105, 491-495.

[36] Koller, M., Stahel, W.A. (2011), Sharpening Wald-Type Inference in Robust Regression for Small Samples, Computational Statistics and Data Analysis 55, 2504-2515.

[37] Lambsdorff, J.G. (2005), The Methodology of the 2005 Corruption Perceptions Index, Transparency International and University of Passau.

[38] MacKinnon, J.G., White, H. (1985), Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties, Journal of Econometrics 29, 305-325.

[39] Murphy, K.P. (2012), Machine Learning: A Probabilistic Perspective, Cambridge (Mass.): MIT Press.

[40] Norris, P. (2001), Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide, Communication, Society and Politics, Cambridge Univ. Press.

[41] O'Mahony, M., Vecchi, M. (2005), Quantifying the Impact of ICT Capital on Output Growth: A Heterogeneous Dynamic Panel Approach, Economica 72, 615-633.

[42] OECD (2001), Understanding the Digital Divide, OECD Publishing, Paris, OECD Digital Economy Papers, 38.

[43] OECD (2011), OECD Guide to Measuring the Information Society 2011, OECD Publishing, Paris.

[44] Pohjola, M. (2003), The Adoption and Diffusion of Information and Communication Technology Across Countries: Patterns and Determinants, New Economy Handbook, 77-100.

[45] Renaud, O., Victoria-Feser, M.-P. (2010), A Robust Coefficient of Determination for Regression, Journal of Statistical Planning and Inference 140, 1852-1862.

[46] Qian, W., Yang, Y. (2013), Model Selection via Standard Error Adjusted Adaptive Lasso, Annals of the Institute of Statistical Mathematics 65, 295-318.

[47] Sala-i-Martin, X., Doppelhofer, G., Miller, R.I. (2004), Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach, American Economic Review 94, 813-835.

[48] Schneider, U., Wagner, M. (2011), Catching Growth Determinants with the Adaptive Lasso, German Economic Review 13, 71-85.

[49] Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society Series B 58, 267-288.

[50] Tibshirani, R.J. (2013), The Lasso Problem and Uniqueness, Electronic Journal of Statistics 7, 1456-1490.

[51] van Ark, B., O'Mahony, M., Timmer, M.P. (2008), The Productivity Gap between Europe and the United States: Trends and Causes, Journal of Economic Perspectives 22, 25-44.

[52] Varian, H.R. (2014), Big Data: New Tricks for Econometrics, Journal of Economic Perspectives 28, 3-28.

[53] Wood, S.N. (2001), mgcv: GAMs and Generalized Ridge Regression for R, R News 1, 20-25.

[54] Wood, S.N. (2006), Generalized Additive Models: An Introduction with R, Boca Raton: Chapman & Hall/CRC.

[55] Wunnava, P.V., Leiter, D.B. (2009), Determinants of Intercountry Internet Diffusion Rates, American Journal of Economics and Sociology 68, 413-426.

[56] Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association 101, 1418-1429.

[57] Zou, H., Hastie, T. (2005), Regularization and Variable Selection via the Elastic Net, Journal of the Royal Statistical Society Series B 67, 301-320.