**Open access is not a panacea, even if it’s radical – an empirical study on the role of shadow libraries in closing the knowledge access inequalities**

**Balazs Bodo (University of Amsterdam), Daniel Antal (CEEMID), Zoltan Puha (Tilburg University)[[1]](#footnote-1)**

**Abstract**

Library Genesis is one of the oldest and largest illegal scholarly book collections online. In the middle of 2018, this shadow library (Joe Karaganis 2018) is hosting and making available without the authorization of copyright holders more than 2 million scholarly publications, monographs, textbooks, as well many works of fiction. It also provides a back-up archive of scientific papers illegally accessed via the Sci-Hub service. Using new, extensive, first hand direct observation data on its usage, this paper follows up on earlier studies (Bodó 2018a; Cabanac 2015) on the role this shadow library plays in the global scholarly publication ecosystem.

The paper analyzes a set of weblogs of one of the Library Genesis mirrors, provided to us by one of the administrators of the service. The weblogs contain records of individual book downloads from the period between September 2014 and March 2015. We use the date, the book identifier, and the geo-coordinates of the downloader included in the dataset to reconstruct the global black-market demand for scholarly literature. We then proceed to build a model to explain this traffic with various macroeconomic indicators on the global stage, and with economic, educational, R&D, and other cultural consumption indicators on NUTS-2 level in the European Union.

# Introduction

Library Genesis (LG or LibGen) is a copyright infringing online collection of scholarly works. It consists of an online catalogue of an ever-growing list of works[[2]](#footnote-2), accessible through a simple web interface. The catalogues itself contains bibliographic metadata as well as various file identifiers[[3]](#footnote-3) which allow users to search and download the digital versions of books from various third-party services, including centralized services, and decentralized p2p networks. All elements of the LG web service are freely available for anyone to download, including the webserver code, the most current copy of the database, or the works themselves.

LibGen contains several collections. Its main focus is scholarly works: scientific monographs, edited volumes, textbooks, handbooks. It also serves as a repository for scientific articles downloaded by the users of SciHub, another copyright infringing shadow library, focusing solely on journal articles (Bohannon 2016), and a separate catalogue of literary works and comics.

The legal status of LibGen is understood to be copyright infringing both by the rights holder, authors, users, and operators of the service (Bodó 2018b). A New Your court issued a default judgement against the sites SciHub, Library Genesis, and their operators, finding them liable for willful copyright infringement, ordering Alexandra Elbakyan, the operator of SciHub, and the anonymous operators of LibGen to pay damages of $15M, and confiscating the domain names (Elsevier Inc. et al v. Sci-Hub et al, Case No. 1:15-cv-04282-RW 2017). A Virginia district court ordered the domains names to be blocked in the US (American Chemical Society v. Sci-Hub d/b/a www.sci-hub.cc, John Doe 1-99 2017). The domain names of the services were at certain points in time blocked in Russia, where these services are thought to be located (Dalmeet Singh Chawla 2018), and by a number of ISPs in Europe. Online service providers, such as Facebook, have also been filtering links to LibGen.

The administrators of LibGen (and for that matter, SciHub) are not contesting the legal assessment (Bodó 2018b; Elbakyan 2015). Neither do those who use these services by downloading or uploading materials from / to them (Swartz 2008). Yet, there seems to be a widely shared (but certainly not universal (Barczak 2017)) consensus among many in the academic sector, about the moral acceptability of such radical open access practices (Bodó 2016; Barok et al. 2015; Gardner and Gardner 2017; Travis 2016; Taylor 2006). Willful copyright infringement in the research and education sector is seen as an act of civil disobedience to resist the business models in academic publishing which have faced substantial criticism in recent years for unsustainable prices and outstanding profit margins they charge for works produced, and reviewed for free by the scholarly community (Mars and Medak 2015). Since shadow libraries are the products of the cooperation of scholars themselves, who contribute texts, and other resources (such as donations, volunteer work, etc.), shadow libraries represent a ‘bottom-up’ approach to open access, a physical approximation of the platonic ideal of knowledge sharing in academia, that would exist if there were no legal, economic, or institutional barriers to the circulation of scholarly knowledge.

The problematic nature of the current economic organization of scholarly publishing has long been acknowledged (Swartz 2008; Suber 2013; Armstrong et al. 2010; Bruijns et al. 2017; Davis and Walters 2011; Holdren 2013; Krikorian and Kapczynski 2010). The traditional model of academic publishing relies on access control, where publishers sell steeply priced subscriptions to journals, books to libraries, and textbooks to students. Its alternative, open access publishing shifts the costs from readers to authors by charging article processing prices for authors in exchange for free open access of the published articles.[[4]](#footnote-4) Both of these business models are exclusionary in one form or another. Access control regimes affected the least resourceful institutions first, but in recent years even the most well-endowed US Ivy league universities warned about the financial unsustainability of subscription fees (Faculty Advisory Council 2012), or cancelled some contracts with journal publishers (Gaind 2019). On the other hand, the article processing fees associated with the now standard Golden Open Access regimes create publication barriers for researchers without institutional budget to cover such fees. In recent years multiple institutional consortia and national science agencies in charge of the agreements with academic publishers let their agreements lapse in hope to reach a financially more sustainable deal. (Kwon 2018; Unit 2019; Max Planck Society 2018; Hungarian Academy of Sciences Electronic Information National Programme Service 2018; UC Office of the President 2019)

Shadow libraries such as Library Genesis and SciHub were created in response to the complex institutional, political, financial, economic conditions that limited access to knowledge at the geographic, and institutional peripheries of academia (Bodó 2018b; Swartz 2008; Bodó 2016). However, since these services are now deeply embedded in the current system of circulation of scholarly knowledge (Himmelstein et al. 2018; Bohannon 2016), their current use is probably more complex than simply serving disadvantaged scholars, low income countries, or underfinanced institutions.

There are very few empirical studies on the extent and potential impact of book piracy in general, and scholarly piracy in particular. There are many possible explanations why online book piracy was rarely in the headlines: e-book markets and audiences are still relatively small compared to print; electronic reading device penetration is much lower than music players and the likes; and print is probably still a preferred format for many. Yet, while e-book piracy is definitely present (as noted by USTR 301 reports, and various surveys), its volume and economic value is perceived to low, especially compared to the losses suffered by the music and audiovisual sectors (Poort et al. 2018). E-book black markets failed to develop their own Napster service, and book piracy sites remained local, fragmented and marginal. As a result, it remained to be hard to study the supply and demand through these illicit services. The few existing studies in the general e-book piracy space, such as (Camarero, Antón, and Rodríguez 2014) and (Reimers 2014) echo findings of studies on music and audiovisual piracy: displacement effects are detrimental mostly for best sellers, long tail content enjoys discovery effect, and individual propensity to pirate depends upon individual norms and attitudes, peer pressure, price sensitivity and technical expertise. In general, however, only a very small segment of the population is involved in e-book piracy.

The high profile investigation against, and the related suicide of Aaron Swartz, the author of the Guerilla Open Access Manifesto (Swartz 2008), and the open rebellion of Alexandra Elbakyan (Elbakyan 2016), the administrator of SciHub brought the issue of scholarly piracy into the mainstream, and resulted in a number of empirical studies on this phenomenon. The research was also aided by the openly accessible LibGen catalogue, and the dataset on SciHub usage released by Elbakyan in 2016 (Elbakyan and Bohannon 2016). Cabanac (2015) offers a rudimentary analysis of the LibGen Catalog, while Greshake does the same for the SciHub dataset (2017). Bodo (2018a) uses a download dataset of LibGen usage from 2012 and finds that the most popular titles in LibGen are widely available via Amazon in various print formats, suggesting that the library’s main role is not the distribution of titles inaccessible via legal alternatives. The study, however, also found that cheap and easy electronic availability (both individual, and institutional) was limited at least in 2013-14, and downloaded works tended to be significantly more expensive that those which haven’t been downloaded. The issue of e-book availability, and its potentially negative impact on the ability of libraries to serve their patrons was also confirmed in a more recent study by Giblin at al. (2019).

Himmelstein (2018) analyzed the SciHub catalog, and found that in many scientific domains it offers more comprehensive access to pay-walled articles that even the best US academic libraries. Muller and Iriarte (2017) measured the availability and access of journal articles cited by university of Geneva researchers in 2015-16 via various sources including SciHub, and found that compared to legal availability, piratical access plays very little role. This is in line with a number of studies from multiple scientific disciplines, which found that the overall weight and impact of this piratical access channel remained marginal (Timus and Zakaria 2016; Babutsidze 2016; Androcec 2017)

Regarding the geographic usage of shadow libraries both Bodo (2018a) analyzing LibGen, and Bohannon (2016) analyzing SciHub data agree that the these services are widely used in both developed and developing countries. This fact suggests the existence of multiple, separate logics that produce the use of scholarly piracy. In rich, North American, and Western European countries, users turn to SciHub and other similar venues probably for convenience (Gardner and Gardner 2017). On the other hand, studies from developing countries suggest a substantial access problem in the Global South, which may drive scholarly piracy (Machin-Mastromatteo, Uribe-Tirado, and Romero-Ortiz 2016; Bruijns et al. 2017; Corrales-Reyes 2017).

In this paper we use a large dataset of directly observed downloads from one of LibGen’s mirror sites. We use this dataset to model what kind of macroeconomic, and institutional conditions may explain the use of shadow libraries. We are particularly interested in the potential function of shadow libraries to mitigate income-related access problems at the peripheries. We are testing the following two hypotheses:

*H1: Globally, per capita shadow library usage is more prominent in lower income countries, controlling for internet penetration.*

We also test the same hypothesis within the European Union, where a much richer dataset allows us to conduct analysis on significantly smaller, sub-national statistical units.

*H2: Within the European Union, the use of shadow libraries is more prominent in lower income EU regions, controlling for the number of academics in the region.*

In addition, in Europe the richness of additional data sources allowed us to test if there are other, less intuitive spatial or social patterns, which could offer more detailed insight into scholarly piracy. We compiled a rich dataset from various European official data sources, such as EUROSTAT and Eurobarometer, and used various modelling techniques, such as random forest simulation to identify and test additional explanatory variables which we could integrate into our piracy models.

# Data overview & descriptive statistics

Multiple sites offer access to books in the LibGen database. The dataset we analyze herein is a weblog of one of such LibGen mirror sites, which has been in continuous operation since at least 2012. The data was provided to us by an anonymous administrator through private correspondence during 2015. Each record in the dataset contained a timestamp, a unique document ID from the LibGen catalogue, and an IP address. We converted IP addresses to Geolocation data using Maxmind[[5]](#footnote-5)’s GeoIP database, and discarded the IP addresses. After the removal of obvious bot traffic[[6]](#footnote-6), and traffic from known TOR exit servers, the logs contained 16.133.680 records over a period of 135 days from between 09/27/2014 and 03/01/2015. Figure 1 shows the daily number of downloads. Except for two periods with no data, the logs raise no apparent doubts about the validity of the information within.

Figure 1: Daily aggregate download volumes

In May 2015, at the end of the observed period the LibGen database contained little more than 1.6 million records. The weblogs referred to the download of 760.868 books from the LibGen catalogue. Compared to the data from 3 years earlier from the same source (Bodó 2018a), the catalogue grew by half a million records from ~836 thousand to ~1.300.000, while the average daily download volume grew more than threefold from ~41.000 downloads /day to ~136.000 downloads / day.

In Table 1 we listed the first 20 countries by absolute download volume. In the last two columns we listed average daily download per million inhabitants, and the rank of the country by per capita downloads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Country  name | Total  dowloads | download per DAY  PER million | per capita  download rank |
| 1 | United States | 1683353 | 39 | 49 |
| 2 | India | 1272124 | 7 | 131 |
| 3 | Germany | 765170 | 69 | 19 |
| 4 | United Kingdom | 594925 | 68 | 21 |
| 5 | China | 580808 | 3 | 158 |
| 6 | Iran, Islamic Republic of | 563798 | 53 | 35 |
| 7 | Italy | 469676 | 57 | 30 |
| 8 | Canada | 369962 | 77 | 17 |
| 9 | Indonesia | 341269 | 10 | 119 |
| 10 | Spain | 327326 | 52 | 37 |
| 11 | Turkey | 323204 | 30 | 63 |
| 12 | Brazil | 307376 | 11 | 112 |
| 13 | France | 290734 | 32 | 59 |
| 14 | Greece | 237657 | 163 | 3 |
| 15 | Mexico | 200792 | 12 | 108 |
| 16 | Australia | 200109 | 62 | 24 |
| 17 | Russian Federation | 196087 | 10 | 118 |
| 18 | Netherlands | 189747 | 83 | 14 |
| 19 | Vietnam | 179758 | 14 | 101 |
| 20 | Egypt | 169421 | 14 | 102 |

Table 1: Country level statistics for the first 20 countries by aggregate download volume

Looking at the geographic location of downloads one can observe that while most downloads cluster around large urban centers, and locations that coincide with institutes of research and higher education, a substantial amount of activity originates from outside of these intuitive download locations.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 2:Geographical distribution of download locations aggregated over the total observation period.

The content-wise analysis of downloaded works (not reported here) also supports the self-professed claims of LibGen, that it is a predominantly scholarly library used to disseminate academic works indiscriminately across the globe to scholarly communities and individuals interested in learning.

# Global models

Our first efforts try to explain the global per capita download volumes by macroeconomic indicators, such as Population (Total), GDP per capita, PPP (current international $) and internet penetration (Fixed broadband Internet subscribers). We then try to add variables related education and research, such as Literacy rate (adult total , % of people ages 15 and above), and School enrollment, tertiary (% gross), Research and development expenditure (% of GDP), Government expenditure per student, tertiary (% of GDP per capita), scholarly research impact, as measured by the aggregate h-index of the country, macro-statistics from the World Bank, and OCDB statistical databases.[[7]](#footnote-7)

If we plot the number of downloads per population per country (colored by continent), we see that there is substantial variation among countries, and between countries of different continents.

|  |  |
| --- | --- |
|  |  |

Figure 3: Country-level and regional variance of the dependent variable

In the first model, we use the following specification:

We tested this model both as a linear model and using a Poisson regression. Since our dependent variable is a count data, the use of Poisson regression is justified. On the other hand, a negative binomial distribution, does not suit this problem well, therefore we omitted that approach. The outputs of the model can be seen in column (1) of Table 2:

───────────────────────────────────────────────────────────────────────────

DV: dl\_per\_pop DV: dl\_per\_pop\_round

Model 1 Model 2 Model 3

─────────────────────────────────────────────────

(Intercept) -5.26e+03     2.5 \*\*\*    2.5 \*\*

(3.22e+03)    (0.0188)    (0.893)

log(pop\_per\_mil) -83.1          -0.0181 \*\*\* -0.0181

(125)            (0.000537)  (0.0255)

log(gdp) 712            0.531 \*\*\*  0.531 \*\*\*

(376)            (0.00205)   (0.0972)

internet\_per\_pop 1.95e+04 \*\*\* 2.82 \*\*\*   2.82 \*\*\*

(3.36e+03)    (0.0134)    (0.635)

─────────────────────────────────────────────────

N 190            190          190

Null deviance 4.51e+09     9.26e+05   9.26e+05

res.deviance 2.48e+09     3.86e+05   3.86e+05

────────────────────────────────────────────────────────────────────

\*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05.

Table 2: Global models I. (DV: download per capita, N=181)

In the general linear model only the Internet Penetration and the GDP have significant effects (the latter only at a 90% level), both positive. While there is a 0.75 correlation between GDP and Internet penetration, the VIF values of the model show that the model does not suffer from multicollinearity.

With the Poisson regression in column (2) of Table 2, all the variables are highly significant. The VIF values are <2, so we should not worry about multicollinearity here either. The signs of the coefficients are the same as with the linear model: countries with higher gross income, and better internet access download more. Population enters as a highly significant explanatory variable with a negative sign, which may be the result of two factors. On the one hand the knowledge demand of populous countries like China, India, Indonesia are not best served by a predominantly English language shadow library. On the other hand, it is possible that the share of population working in knowledge intensive domains of society do not scale linearly with population.

One possible downside of the Poisson regression is that it cannot deal with overdispersion (only one parameter is estimated). This can lead to underestimated standard errors, which we tested with a Wald test. Since the scale factor in the Poisson model is much higher than 1 (residual\_deviance / df = 385842 /186), we corrected for overdispersion by using a QuasiPoisson regression model, presented in column (3) of Table 2. In this last model GDP and internet penetration are highly significant, and have positive effects.

Taken together, these models suggest a result which contradicts our hypothesis that low(er) income countries may use shadow libraries more to compensate for infrastructural, and funding limitations.

To explore further, we added a number of macroeconomic variables related to tertiary education and research activities. We queried [gross tertiary education enrollment ratio](https://data.worldbank.org/indicator/SE.TER.ENRR), the expenditure on tertiary education per student and the percentage of GDP spending on R&D from 2015 from the World Bank Open Data dataset. We also included the H index of countries from 2015. Due to missing data the sample size was reduced from 235 to 125. For this model and all the following ones, we only include the results of the QuasiPoisson regression as this is the best fit for our data. The results are summarized in Table 3.

────────────────────────────────────────────────────────────────────

Model 4 Model 5 Model 6

─────────────────────────────────────────

(Intercept) 2.44      8.15 \*\*\*  7.12 \*\*\*

(1.67)     (0.267)    (0.358)

log(pop\_per\_mil) -0.0926    -0.326 \*\*\* -0.341 \*\*\*

(0.0875)   (0.0719)   (0.0672)

log(gdp) 0.583 \*\*

(0.181)

internet\_per\_pop 1.61

(1.43)

tertiary 0.0023    0.0102 \*  0.0297 \*\*\*

(0.0043)   (0.00444)  (0.0061)

exp\_tertiary\_pstudent -1.37e-05  9.11e-06  -2.49e-06

(1.57e-05) (1.78e-05) (1.56e-05)

rd 0.0875    0.148     1.56 \*\*\*

(0.103)    (0.11)     (0.323)

h\_index -0.000156  0.000743  0.000896 \*

(0.000491) (0.000446) (0.000447)

tertiary:rd                   -0.0211 \*\*\*

                  (0.00476)

─────────────────────────────────────────

N 86         86         86

Null deviance 4.7e+05   4.7e+05   4.7e+05

res.deviance 1.97e+05  2.53e+05  1.91e+05

────────────────────────────────────────────────────────────────────

\*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05.

Table 3: Global models II. (DV: download per capita)

Including these extra variables in the original model (Table 3 Model 4.) did not produce significant new insights, as the GDP variable seem to capture much of the effect of higher education and research investment. In model (5) we excluded the GDP and internet penetration variables. In the resulting, weaker model, the share of population with tertiary education, and the h-index variable became significant (the latter only at a 90% level), with intuitive results: a larger share of highly educated people, and more relevant scientific output results in higher shadow library use. Interestingly, the share of active tertiary education students, a potential source of shadow library traffic, was not significant. In the last model we introduced a new interaction term between the share of population with tertiary education and RD expenditure, because these two activities are two most obvious sources of shadow library use, but they may be independent from each other. In Model (6) both variables turn highly significant, suggesting that both the size of highly educated population and the RD activities contribute positively to shadow library demand, albeit at a diminishing rate, as the interaction term has a negative sign. We found no effect of public spending on tertiary students, which could have differentiated between countries with publicly and privately funded higher education systems. These models refine the effect of GDP, and identify research and tertiary education as major drivers of shadow library use.

Lastly, we explored the effect of regional differences, because the descriptives suggest (see Figure 3) that there are substantial regional differences. Our last global model is a varying intercept and slope (random effects) model. The model contains the intercept, and the GDP varying with the continent, while the effect of the population and internet penetration is fixed. The DV is download per capita.

The coefficients are the following:

log(1 + gdp\_scaled) (Intercept) pop\_per\_mil\_scaled internet\_per\_pop\_scaled

Africa 1.1025678 8.248046 -0.3200811 0.2780263

Americas 0.2314024 7.858342 -0.3200811 0.2780263

Asia 0.5302122 7.981479 -0.3200811 0.2780263

Europe 0.1876811 8.645233 -0.3200811 0.2780263

North America 0.3898888 8.455814 -0.3200811 0.2780263

Oceania 1.2764129 7.672324 -0.3200811 0.2780263

Table 4: Global models III. Random effects model by continent (DV: download per capita)

From the table above the varying intercept points to higher European and North American download baselines. More interesting is the huge difference in the effect of GDP. The impact of gross income on downloading is much higher for countries in the African continent than for example in Europe.

It is possible that the cause of these differences is geographical in nature, because, for example, shadow library related practices propagate via physical proximity and close trust relationship of individuals. While this may be the case, it is hard to test that hypothesis with the current data. On the other hand, geographic location may also be a proxy for the level of development, and in that case, we can conduct a similar analysis, using the World Bank’s income categories instead of geographic location, and test for the effect of GDP, R&D and educated population in different income categories[[8]](#footnote-8).

Table 5 and 6 shows the outcome of these models.

log(1 + gdp\_scaled) (Intercept) pop\_per\_mil\_scaled internet\_per\_pop\_scaled

High income -0.003871827 0.30688303 -0.1200297 0.3071157

Upper middle income 0.105214527 -0.06604125 -0.1200297 0.3071157

Lower middle income 0.397072298 -0.07077302 -0.1200297 0.3071157

Low income 0.348602516 -0.17360155 -0.1200297 0.3071157

Table 5: Global models IV. GDP random effects model by income category. (DV: download per capita rounded, quasipoission)

tertiary\_scaled rd\_scaled (Intercept) pop\_per\_mil\_scaled internet\_per\_pop\_scaled

High income 0.010932241 -0.05358038 8.543931 -0.2996871 0.4264995

Upper middle income -0.001022486 0.35217880 8.230012 -0.2996871 0.4264995

Lower middle income 0.050455057 1.63851738 8.360397 -0.2996871 0.4264995

Low income 2.797771598 1.45719062 9.857220 -0.2996871 0.4264995

Table 6: Global models V. R&D and education random effects model by income category (DV: download per capita rounded, quasipoission)

Both models point to interesting findings. The effect of GDP is very different in the four income categories: In low income countrues increasing GDP causes much larger shadow library use than in high income countries. The model in Table 6 suggests that in low income countries, extra investment in tertiary education, and R&D activities generates relatively much larger shadow library usage than similar investment in high income countries. In the latter group extra investment into R&D and tertiary education is associated with relatively lower download volumes, while in low income countries the effect is exactly the opposite: higher investment into knowledge intensive social activities generates more demand for black market knowledge. The reason for that is straightforward. In high GDP countries, extra money spent on knowledge intensive activities is more likely to include spending on infrastructures of legal access, lessening the need for grey market venues. On the other hand, in low income countries, where the infrastructure is probably the most lacking, any step to increase the knowledge intensive domains and knowledge-hungry populations is likely to hit infrastructural constraints, leaving some of the demand at the mercy of access provided by shadow libraries.

At first sight, the global models did not support our first hypothesis, that countries with less resources to spend on research and higher education would be more intense users of shadow libraries to offset their infrastructural limitations. On the contrary, our early findings suggested that as countries’ GDP per capita, tertiary enrollment, or research expenditure grows, they also shadow library more intensely. At the aggregation level of individual countries this is hardly surprising: access to knowledge is only one element in the complex infrastructural mix which then produces demand for the knowledge shadow libraries may offer.

We arrive to a more nuanced conditions, if we start to disaggregate the impact of gross income through finding various proxies of general development. The intuition behind this approach is that investment into knowledge intensive societal domains, like R&D and higher education serve different purposes and have different effects at different stages of development. In low income countries higher investment may lead to fast growing knowledge absorption capacity, which may not be met with appropriate infrastructural support. This means that low income countries generate less shadow library usage in general, but within that group, larger investment into knowledge intensive activities have greater and positive impact on usage. In high income countries, the logic is the opposite. While they have larger per capita demand, larger investment in knowledge intensive activities does not further increase black market demand. On the contrary, since extra investment most probably creates better infrastructural conditions, rather than extra knowledge absorption capacity, larger investment leads to relatively lower black market demand.

In summary, having access to a virtually unconstrained source of knowledge may not automatically be able to deliver all the potential benefits. The impact of higher income manifests itself in two forms. On the one hand, it creates knowledge demand through the prominence of knowledge intensive institutions, and knowledge demanding social strata. While higher income certainly expands the knowledge absorption capacity of countries, it may not establish, at the same pace, the adequate institutional frameworks to service that demand. The infrastructure of legal access may be lagging behind the growth of this demand, creating the conditions for shadow library use. It seems that only at higher income levels the extra investment in knowledge intensive domains is able to create the adequate access infrastructures which ultimately moderate shadow library use.

Our hypothesis, in its original form, is only supported among high income countries. But the larger impact of R&D and educational investment on downloads in low income countries also lends support to this hypothesis, albeit is a slightly different form. In low income countries the per capita download starts at a lower base, because the knowledge absorption capacity of these countries is limited, but any extra income produces growth in that absorption capacity, which in turn creates comparatively larger demand in the shadow libraries.

At this stage we should point to some of the limitations in our data, which may affect these findings. First, data may be skewed by the use of VPNs by users whose ISP blocks access to LibGen. Second, in countries with low bandwidth local copies of LibGen, and shadow libraries may serve much of the demand, which produce download figures that are lower than the actual use of this resource.

# The European models

In this section we focus only on downloads form within the European Union. This allows us to address many limitations of the global models. First, we can zoom into regional levels. Geolocation tends to work better in Europe, so we can be more confident of the geographic location of a particular download, and allow us to zoom into sub-national socio-economic units[[9]](#footnote-9). Second, due to a number of datasets produced by Eurostat, we have a much better choice of institutional, economic, and attitudinal variables we can include in our models. Third, the authors being from the Union, we have the advantage of the home field to interpret results.

The European Union is the world’s largest harmonized statistical data collection area with four levels of statistical aggregation starting from national (NUTS0) level to very small territorial units down to NUTS3 level. We selected the NUTS2 regions of the European Union for our environmental analysis. The NUTS2 regions were created for socio-economic statistical purposes and they are designed to maximize intra-unit homogeneity. While NUTS2 social and economic data is not always complete, partly because NUTS2 boundaries change relatively frequently, usually we can still work with 140-260 territories. The Eurobarometer or the European Social Survey is designed to represent NUTS0 (country) levels but apart from larger countries can be re-aggregated at NUTS2 levels with relatively little bias. The NUTS2 level is a good compromise between NUTS0 (country) and the much smaller NUTS3 levels. While NUTS3 levels would allow the comparison of about a thousand environments, we have far less data available on NUTS3 level. Furthermore, on NUTS3 level we would need to tackle problems of non-normal distribution, as on the NUTS3 level our data becomes asymmetric. Therefore, we aggregated the download data over the NUTS2 boundaries, joined them with environmental data, imputed missing variables, and normalized the data.[[10]](#footnote-10) These processes are described in more detail in the supplementary material on methods.

For the European models we used the following data from Eurostat[[11]](#footnote-11), and the Eurobarometer[[12]](#footnote-12) (we included a detailed list of all variables in Appendix 1):

* Size of the environment, such as land area of the NUT2 areas, or the population groups such as total adults, total inhabitants, total researchers per NUT2 region.
* Description of the socio-economic environment on a macro level such as GDP, household income of the NUTS2 area, or research, innovation spending within the area that affect more closely to the research and academic tuition environment.
* Computer use data, which is derived from various European surveys about home, work, school access to broadband internet, use of the internet for banking, purchase of paid book and other content, reading news, or the opposite, the size of population who has no access to computers or the internet.
* Cultural access and participation variable data, such as visiting a public library at least once a year, reading a book at least once a year, and not visiting public libraries more often because of perceived low-quality local supply.

The richness of the European dataset allowed us to purse a deductive modelling strategy, and in addition use advanced statistical methods to explore new patterns in the data.

## Hypothesis testing

First, we test our original hypothesis on the European regional data, namely that

*H2: Within the European Union, the use of shadow libraries is more prominent in lower income EU regions, controlling for the number of academics in the region.*

Historical accounts, which reconstruct the development and raison d'être of shadow libraries (Bodó 2018b; 2011; Bodó and Lakatos 2012; Bodó 2016) suggest that inadequate legal access alternatives may be the main drivers of digital piracy in this region in general. The authors, who originally had close relationship with academia in the region, also have extensive personal experiences about the lacking infrastructural conditions of scholarly work, and the extensive use of piratical resources to provide competitive higher education degrees for students with an eye on the European job market, and produce research relevant in the European and global arena. Our first-hand experience matches with other accounts from the economic, academic periphery: shadow libraries may offer a way to overcome income related infrastructural limitations for scholars.

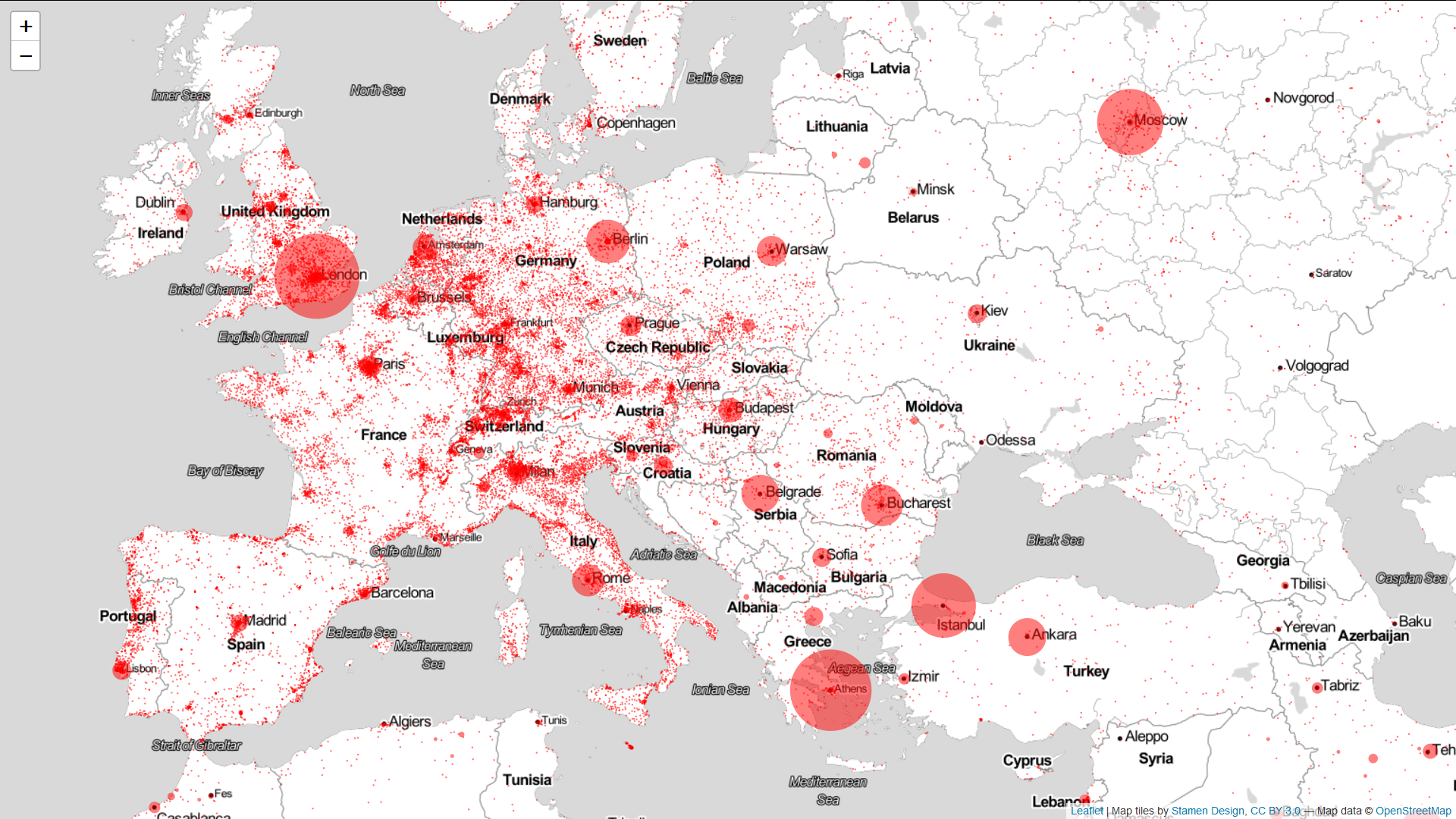


Figure 4: European download locations

We chose download per capita as a dependent variable for our model, with the following specification

Where *GDP\_PPS* is the price-adjusted version of the GDP indicator, using purchasing power standards rather than Euros to account for the differences is purchasing power (used in a logarithmical form); *Researcher\_employment\_pct* is the percentage of R&D personnel and researchers in the workforce, and *Internet\_banking\_pct* is the percentage who used the internet for online banking in the population. We treat this latter variable as a rough proxy for internet proficiency.

This model is somewhat comparable to the global model, as it interrogates the same underlying dynamics, albeit with variables that better approximate the factors in question. Instead of R&D expenditure[[13]](#footnote-13) of the global model we can use the share of researchers in the local workforce, and instead of the internet penetration, we have data on the advanced use of the internet.

As before, we use a QuasiPoisson regression model to correct for overdispersion, and account for the fact that we model count data. The VIF values of the regression are all <2, so we do not have to worry about multicollinearity.

──────────────────────────────────────────────────────────────────────────────────────────────────────────────

download per capita

Model 7 Model 8 Model 9 Model 10 Model 11

─────────────────────────────────────────────────────────────────────────────────────────

(Intercept) 6.438 \*\*\* 6.295 \*\*\* -0.167     6.353 \*\*\* 4.050 \*\*\*

(0.794)    (0.838)    (2.468)    (0.802)    (1.110)

log(gdp\_pps) 0.247 \*\*  0.242 \*\*  0.175 \*   0.258 \*\*  0.490 \*\*\*

(0.077)    (0.081)    (0.075)    (0.078)    (0.105)

researcher\_employm 0.697 \*\*\* 0.683 \*\*\* 0.515 \*\*\* 0.702 \*\*\*

ent\_pct (0.057)    (0.063)    (0.097)    (0.055)

internet\_use\_banki -0.011 \*\*\*          -0.017 \*\*\* -0.009 \*\*  -0.003

ng\_pc (0.003)             (0.004)    (0.003)    (0.005)

internet\_purchases          -0.006

\_last\_year\_pc (0.003)

log(disposable\_inc                   0.783 \*\*

ome) (0.281)

edu\_attainment\_tot                   0.007

al (0.009)

gerd                            -0.070     0.059

                           (0.053)    (0.076)

───────────────────────────────────────────────────────────────────────────────────────────

null.deviance 2990524.371     2990524.371     2990524.371     2990524.371     2990524.371

res. deviance 1415805.433     1507337.393     1337655.185     1396838.896     2455507.137

─────────────────────────────────────────────────────────────────────────────────────────────────────────────

\*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05.

Table 7: European models I. (DV: download per capita)

In Model 7, our base model, all independent variables are highly significant, and we explain ~72% of the variance. The per capita downloads grow with the GDP, as well as with the share of researchers in the workforce. On the other hand, shadow library usage is moderated by internet proficiency.

The interpretation of the former two effects is straightforward, and is in line with the findings of our global models. Shadow library usage is positively correlated with income. It is also intuitive that the researcher population drives shadow library demand. The negative sign of internet proficiency variable, however, demands some explanation. That variable can be a proxy of many different online skills: a better knowledge of digital piracy, including the use of shadow libraries; skills to use the internet for online purchases; and skills to the hide the traces of illicit activities, via the use of Virtual Private Networks, and TOR browsing.[[14]](#footnote-14)

To explore further what the online banking variable may refer to, we replaced it with the percentage of population who used the internet for online shopping in Model 8. The intuitive assumption here was that a negative relationship (a replacement effect) exists between online shopping and digital piracy. Though the sign of the variable was negative, the relationship was not significant. Since the online shopping and the online banking variables are highly correlated (Pearsons’s: 0.86, p < .001), we can assume that online banking already captures some of that effect.

In Model 9, we tested further variables, such as the effect of disposable income, and the share of population with tertiary education. The effect of disposable income is positive and significant at a 95% level, while the effect of education is nonsignificant. While this model suffers from higher multicollinearity, it is clear that the individual income effect and the macro-income indicator both point to the same direction: people download more from more affluent regions.

In models 10 and 11 we introduced the R&D variable, but found no statistically significant effect.

That being said, the R&D expenditure (gerd) becomes significant, and with a negative sign, if the dependent variable is the downloads per researchers (see Table 7, Model 12). While the effect signs for the other relevant variables (GDP\_PPS, internet proficiency) remain the same, when we normalize downloads for the number of researchers, a higher R&D spending has a moderating effect on the per researcher downloads. This may be the first sign which points to a structural link between the amount of investment into knowledge infrastructures, and scholarly piracy. We should note that the download per researcher models have a much worse fit than the per capita models.

────────────────────────────────────────────────────────────────────────────────────

download per researchers

Model 12 Model 13 Model 14

────────────────────────────────────────────────────

(Intercept) 5.530 \*   8.034 \*\*  6.486 \*\*\*

(2.240)    (2.546)    (1.147)

log(gdp\_pps) 0.161 \*   0.174 \*   0.175

(0.071)    (0.078)    (0.113)

log(disposable\_income) 0.148     -0.143

(0.255)    (0.291)

edu\_attainment\_total 0.008     -0.000

(0.008)    (0.009)

gerd -0.253 \*\*  -0.310 \*\*\* -0.155

(0.079)    (0.088)    (0.901)

internet\_use\_banking\_pc -0.018 \*\*\*

(0.004)

internet\_purchases\_last\_year\_          -0.006

pc

         (0.004)

log(gdp\_pps):gerd                   -0.024

                  (0.084)

────────────────────────────────────────────────────

null.deviance 495798.787     495798.787     495798.787

deviance 362061.621     398536.668     414631.765

────────────────────────────────────────────────────────────────────────────────────

\*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05.

Table 8: European models II. (DV: download per researcher)

In models where the dependent variable is the raw download count (see the supplementary materials), we find results consistent with those above: wealth and researcher population have significant positive effects, internet proficiency has significant negative effects, R&D spending, educational attainment, disposable income, or online shopping variables are not or only weakly (at 95% level) significant.

So far, we have established that income and the researcher population are the most significant positive drivers of shadow library usage in Europe. In the next step, we build a simple model in which these two variables are in interaction. In this model (see Table 8), raw, not normalized download count is the dependent variable, while GDP\_PPS is used is its natural form.

─────────────────────────────────────────────────────────────────────────

Downloads

Model 15

─────────────────────────────────────

(Intercept) 8.546324e+00 \*\*\*

(1.587638e-01)

gdp\_pps 1.050631e-05 \*\*\*

(1.278794e-06)

researcher\_employment\_pct 9.183301e-01 \*\*\*

(1.158645e-01)

gdp\_pps:researcher\_employment\_pct -3.479490e-06 \*\*\*

(7.194871e-07)

─────────────────────────────────────

null.deviance 7.192467e+06

residual deviance 3.556374e+06

─────────────────────────────────────────────────────────────────────────

\*\*\* p < 1.000000e-03; \*\* p < 1.000000e-02; \* p < 5.000000e-02.

Table 9: European models III. (DV: download count)

In Model 15 all coefficients are highly significant, with a negative interaction term. This suggests that within the EU, even if two regions have similar researcher density, high income regions use shadow libraries more. The difference between low and high income regions is significant, and diminishes with the growth of income only slightly.



Figure 5: Interaction effects between GDP\_PPS and researcher employment percentage (DV: download count)

**Discussion**

The European models are in line with our global models and suggest that similar logics are in place within the European Union as globally. We identified two main drivers of the demand of pirated knowledge: the presence of knowledge-intensive economic activity and GDP. Just as in the case of the global models, the number of researchers sets the baseline demand: the production of knowledge requires knowledge. However, the income-related infrastructural limitations do not translate into relatively higher shadow library use, because income also defines a knowledge absorption capacity. We found some support for this in the download per researcher model, where we found a strong, and significant negative effect of R&D investment on per researcher download volumes. In the interaction model we have also seen that some of the extra income probably sustains infrastructures which better cater for the extra demand.

Researchers in low income regions may face many problems, legal access being only one of them. The authors have personal experience of at least some of the hurdles that may limit an intensive use of openly accessible knowledge wealth. Researchers at the economic, academic peripheries may not be able to fully sustain themselves through a single academic research job. The need to do second and third jobs to sustain themselves may limit their time they can dedicate to library use, piratical or otherwise. Also, many of them see less usefulness of a predominantly English language shadow library, if their educational and research activities are not intended for the English-speaking global market.

## Inductive models

So far, we have been replicating the global models within the EU. The dataset, however, allows us to switch modelling approaches, and look for patterns in the data on which new hypotheses can be formulated. First, we looked at the spatial autocorrelation of data, second, we used different approaches to find new patterns in the data.

**4.2.1. Spatial autocorrelation**

The analysis of spatial autocorrelation reveals if shadow library usage is geographically clustered, for example, because the underlying social, economic activities are also clustering, or because user communities are clustered (for linguistic reasons, or because the knowledge about shadow libraries dissipates in close-knit trust networks). The ability to examine spatial autocorrelation is an important check on the robustness of our methodology. Given that we do not have access to individual downloader data, only territorial aggregates of downloads, we want to make sure that downloading in geographical space is not happening randomly. We have examined the spatial autocorrelation using the `spdep` package (Bivand et al. 2019).

In the case of the download count variable, Moran's I statistic takes the value of 0.042 with a p-value of 0.094, so we can only reject the randomness of downloads at a 90% significance level. The positive z value means that the downloads are clustered, i.e. NUTS2 regions with high download numbers tend to be neighbors of NUTS2 regions with high download numbers. If we run the same test on the GDP adjusted by purchasing power, we see a very similar level of spatial autocorrelation: Moran's I statistic is 0.044017, p-value = 0.077.

The results, at least on the NUTS2 level do not point to well-defined download centers within Europe, and their strong similarity with how GDP is geographically distributed, suggests that it is unlikely that downloads follow a random pattern, and are closely related to the social, economic factors that define the wealth of a region in general.

**4.2.2. Random forest models**

We also ran random forest models to identify which variables out of the 50 available may play an important role in explaining downloads. Classical model selection methods, such as backward or forward stepwise regression are generally based on strong assumptions about the functional form of the model or the distribution of residuals. They are particularly sensitive to cases where there is a strong multicollinearity present among the variables, as in our datasets.   The random forest method, was mainly developed to solve classification or regression problems, but it has been long recommended for use for variable (pre-)selection (Molnar 2019).

In a series of models, not reported here, we first narrowed down the basic geographical and demographic forces attracting higher download counts, such as the land area, population and population density, or researcher population density of the regions. We also normalized count with land area, population and researcher count to get a deeper insight into the non-trivial social factors that attract a heavier reliance on the research black market (Gelman and Hill 2007).

In the second round we ran the random forest algorithms on the various forms of the count data to identify the most important social, cultural and economic variables[[15]](#footnote-15). The form of random forest method we used operate by creating random samples of our dataset, and fitting regression trees on these subsets of the dataset. By repeatedly splitting the dataset, and testing a limited number of features at a time, the random forest algorithm usually does not require strict conditions on the residual errors and it is insensitive to multicollinearity. One draw-back of random forest, like many machine learning models, is that it uses its own metrics of accuracy. For comparability, we used a model-agnostic feature importance metric by Molnar et. al. (Molnar 2019), a metric that results in comparable metrics for random forests and any other statistical model. This feature importance algorithm shuffles the values of the predictors, and measures the change in a loss function (in our case, mean average error increase in the targeted count, count\_per\_area and count\_per\_researcher variables) for each shuffle – the larger the increase in mean average error, the more important to use the (correct values) of the predictor.

Since the outcome of these models do not differ substantially from what we have reported already, here we only report the results of the models which take the per capita downloads as the dependent variable, and uses a smaller, but richer dataset, which includes the EUROBAROMETER data next to the EUROSTAT data we have been using so far.[[16]](#footnote-16) For the purposes of the modeling, we scaled the variables to unit variance, so that they have equal weight in the variable selection process.

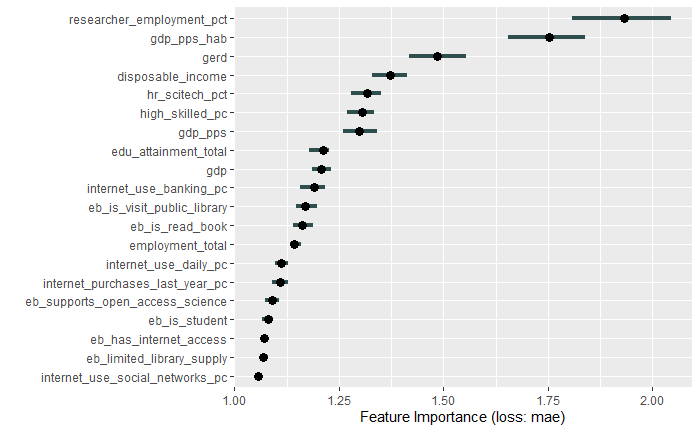


Figure 7: Random Forest feature importance of EUROSTAT+EUROBAROMETER. (DV: count per capita, number of runs: 100)

The feature importance graph (Figure 7) identifies as relevant the same variables we already included in our linear models: the share of researchers in the workforce, GPD per capita in purchasing parity units, R&D investment. In addition, the share of library using, and book reading populations from among the EUROBAROMATER variables are also somewhat relevant.

Subsequently, we have included the newly identified EUROBAROMETER variables into the QuasiPoission regression models, with the per capita, per researcher and raw count as dependent variables.

─────────────────────────────────────────────────────────────────────────────────────────────────────────────

download per capita  download per researcher

Model 16 Model 17a Model 17b

─────────────────────────────────────────────────────────────────────────────────────────────

(Intercept) 6.204 \*\*\* 8.160 \*\*\* 6.905 \*\*\*

(0.843)    (0.813)    (0.886)

log(gdp\_pps) 0.261 \*\*  0.006     0.053

(0.080)    (0.078)    (0.082)

researcher\_employment\_pct 0.673 \*\*\*

(0.056)

eb\_is\_visit\_public\_library -1.116 \*\*  -1.391 \*\*

(0.415)    (0.428)

eb\_limited\_library\_supply                   3.963 \*\*\*

           (0.886)

─────────────────────────────────────────────────────────────────────────────────────────────

null.deviance 2553172.101     350303.560     350303.560

res.deviance 1061645.079     324682.437     312052.421

──────────────────────────────────────────────────────────────────────────────────────────────────────────────

\*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05.

Table 10: European models IV: Eurobarometer variables (DVs: download per capita, download per researcher)

In various model configurations most of the EUROBAROETR variables remained insignificant.[[17]](#footnote-17) That being said, the variable on the share of population who reported to have visited a library in the 12 previous months was highly significant in both the download per capita and the download per researcher models, with a negative sign. Higher library use in the population goes hand in hand with lower pirate library usage. In addition, in the download per researcher model (Model 2b) we found that the more people report not using library because of its inadequate supply[[18]](#footnote-18), the higher the use of shadow libraries. While both these findings support our hypothesis that the quality of legal access infrastructures has strong impact on shadow library usage, we treat these findings with some caution. The usefulness of these EUROBAROMETER variables are relatively limited due to the limited number of respondents, and the reliability of the statistic on a regional level.

# Conclusions

In our earlier work on scholarly piracy, we have conducted a supply side analysis. That work established that a significant chunk of the shadow library supply is not available in digital formats, and a significant share of the downloads concentrate on such legally inaccessible works. This offered a simplistic hypothesis: shadow library usage is mostly driven by market failures and the lack of convenient, digital legal access alternatives.

In our present work we offer a more detailed and elaborate picture on piratical demand of scholarly works. Using similar models to explain global differences in shadow library use on a country level, and a more fine-grained analysis of scholarly piracy within the EU we arrive to similar conclusions.

Scholarly literature is a special information good. It is mainly used as an input for knowledge-intensive social and economic activities: (higher) education, and research and development. Its consumers are almost exclusively highly educated, who possess some online proficiency to access often concealed shadow libraries. For the same reasons they can safely be assumed to be aware of the legal and ethical dilemmas around the illicit access of copyrighted scholarly publishing.

We have found two significant demand drivers of scholarly piracy: GDP and the size of knowledge-intensive sector. Contrary to our initial, somewhat naïve assumption, we have found that gross income and piracy is positively correlated. Free-to access piratical resources are used more in high income territories, with potentially better legal access opportunities, such as libraries, and other institutional and individual access alternatives. This suggests that the lack of legal access infrastructures does not provide a satisfactory explanation for how shadow libraries are used.

In this article we have offered two alternative explanations. First, we have offered a model to differentiate the effect of income on knowledge demand at different levels of economic development. In our global models we have shown that in low income countries extra income has a much greater impact on shadow library demand than in high income countries. This may have to be related to the mechanics of extra spending on knowledge intensive sectors. In low income countries extra spending increases demand, as it expands the scope and amount of potential demand; while in high income countries extra spending may result in better legal supply infrastructures, rather than the further expansion of demand.

Second, our European models suggest there are other, social, economic factors, which limit the capacity to use and absorb freely accessible knowledge in the knowledge intensive sectors of low-income regions. Even if the size of the knowledge intensive sector is comparable to those in richer regions, less affluent regions face constraints, which limit their ability to use and absorb knowledge from freely accessible resources. That being said, we have found some evidence to the importance of good legal access infrastructures: where libraries are used, and found adequate, less scholarly piracy takes place.

These findings can also serve as a warning to the global open access movement gaining momentum. Open access, legal or piratical, is hardly a panacea, and access to knowledge is not the only, and as our study shows, potentially not even the most important hurdle in front of local knowledge intensive social and economic activities. Access is only one aspect which defines the global dissemination, and the local use and usefulness of knowledge. A lot depends on the local conditions which ultimately define how much of the freely accessible knowledge can be absorbed and utilized by local individual and institutional actors.

This study has many limitations. The data it relies on is relatively dated. The geolocation of download data may be inaccurate due to a number of factors: the inaccuracy of IP address-to-geolocation dataset, our inability to fully detect and isolate clandestine traffic via VPNs, and TOR network, and automated traffic via bots and scrapers. We wish we had better datasets to separate different forms of demand: educational uses from university networks, R&D related demand by economic actors, and university research. Hopefully we’ll be able to address these issues in further work.

# References

American Chemical Society v. Sci-Hub d/b/a www.sci-hub.cc, John Doe 1-99. 2017 1–3. United States District Court for the Eastern District of Virginia Alexandria Division.

Androcec, Darko. 2017. “Analysis of Sci-Hub Downloads of Computer Science Papers.” *Acta Univ. Sapientiae, Informatica* 9 (1): 83–96.

Armstrong, C., J. De Beer, D. Kawooya, A. Prabhala, and T. Schonwetter. 2010. *Access to Knowledge in Africa: The Role of Copyright*. Uct Pubns. http://books.google.com/books?hl=en&amp;lr=&amp;id=mzoPtBoQkEcC&amp;oi=fnd&amp;pg=PR5&amp;dq=vera+franz+a2k+mit&amp;ots=EFQceWqCSH&amp;sig=Jirm604ts2dxnSNGpc83\_A7AiZg.

Babutsidze, Zakaria. 2016. “Pirated Economics.”

Barczak, Gloria. 2017. “Academic Piracy.” *Journal of Product Innovation Management* 34 (6): 716–17. https://doi.org/10.1111/jpim.12422.

Barok, Dušan, Josephine Berry, Balázs Bodó, Sean Dockray, Kenneth Goldsmith, Anthony Iles, Lawrence Liang, et al. 2015. “In Solidarity with Library Genesis and Sci-Hub.” Http://Custodians.Online/. 2015. http://custodians.online/.

Bivand, Roger, Micah Altman, Luc Anselin, Renato Assunção, Olaf Berke, Andrew Bernat, Guillaume Blanchet, et al. 2019. *Spdep: Spatial Dependence: Weighting Schemes, Statistics* (version 1.1-3). https://CRAN.R-project.org/package=spdep.

Bodó, Balázs. 2011. “Coda: A Short History of Book Piracy.” In *Media Piracy in Emerging Economies*, edited by Joseph Karaganis. New York: Social Science Research Council.

———. 2016. “Pirates in the Library An Inquiry into the Guerilla Open Access Movement.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2816925.

———. 2018a. “Library Genesis in Numbers: Mapping the Underground Flow of Knowledge.” In *Shadow Libraries - Access to Knowledge in Global Higher Education*, edited by Joe Karaganis. Cambridge, MA: MIT Press.

———. 2018b. “The Genesis of Library Genesis: The Birth of a Global Scholarly Shadow Library.” In *Shadow Libraries - Access to Knowledge in Global Higher Education*, edited by Joe Karaganis. Cambridge, MA: MIT Press.

Bodó, Balázs, and Zoltán Lakatos. 2012. “Theatrical Distribution and P2P Movie Piracy: A Survey of P2P Networks in Hungary Using Transactional Data.” *International Journal of Communication* 6 (1).

Bohannon, John. 2016. “Who’s Downloading Pirated Papers? Everyone.” *Science*, 2016. http://www.sciencemag.org/news/2016/04/whos-downloading-pirated-papers-everyone.

Bruijns, Stevan R, Mmapeladi Maesela, Suniti Sinha, and Megan Banner. 2017. “Poor Access for African Researchers to African Emergency Care Publications: A Cross-Sectional Study.” *Western Journal of Emergency Medicine* 18 (6): 1018–24. https://doi.org/10.5811/westjem.2017.8.34930.

Cabanac, Guillaume. 2015. “Bibliogifts in LibGen? A Study of a Text‐sharing Platform Driven by Biblioleaks and Crowdsourcing.” *Journal of the Association for Information Science and Technology*.

Camarero, Carmen, Carmen Antón, and Javier Rodríguez. 2014. “Technological and Ethical Antecedents of E-Book Piracy and Price Acceptance: Evidence from the Spanish Case.” *The Electronic Library* 32 (4): 542–66.

Corrales-Reyes, Ibraín Enrique. 2017. “Sci-Hub and Evidence-Based Dentistry: An Ethical Dilemma in Cuba.” *J Oral Res* 6 (7): 175. https://doi.org/doi:10.17126/joralres.2017.054.

Dalmeet Singh Chawla. 2018. “Sci-Hub Blocked in Russia Following Ruling by Moscow Court.” Chemistry World. December 3, 2018. https://www.chemistryworld.com/news/sci-hub-blocked-in-russia-following-ruling-by-moscow-court/3009838.article.

Davis, Philip M, and William H Walters. 2011. “The Impact of Free Access to the Scientific Literature: A Review of Recent Research.” *Journal of the Medical Library Association: JMLA* 99 (3): 208.

Elbakyan, Alexandra. 2015. “LETTER Addressed to Judge Robert W. Sweet from Alexandra Elbakyan Re: Clarification of Details. (Ajs).” New York: Elsevier Inc. et al v. Sci-Hub et al, Case No. 1:15-cv-04282-RW.

———. 2016. “Sci-Hub Is a Goal, Changing the System Is a Method.” Engineuring. 2016. https://engineuring.wordpress.com/2016/03/11/sci-hub-is-a-goal-changing-the-system-is-a-method/.

Elsevier Inc. et al v. Sci-Hub et al, Case No. 1:15-cv-04282-RW. 2017. UNITED STATES DISTRICT COURT SOUTHERN DISTRICT OF NEW YORK.

Faculty Advisory Council. 2012. “Faculty Advisory Council Memorandum on Journal Pricing.” Cambridge, MA: Harvard University Library. http://isites.harvard.edu/icb/icb.do?keyword=k77982&tabgroupid=icb.tabgroup143448.

Gaind, Nisha. 2019. “Huge US University Cancels Subscription with Elsevier.” *Nature* 567 (7746): 15–16. https://doi.org/10.1038/d41586-019-00758-x.

Gardner, Carolyn Caffrey, and Gabriel J Gardner. 2017. “Fast and Furious (at Publishers): The Motivations behind Crowdsourced Research Sharing.” *College & Research Libraries* 78 (2).

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge ; New York: Cambridge University Press.

Giblin, Rebecca, Jenny Kennedy, Charlotte Pelletier, Julian Thomas, Kimberlee G Weatherall, and Francois Petitjean. 2019. “What Can 100,000 Books Tell Us about the International Public Library e-Lending Landscape?” *Forthcoming, Information Research*.

Greshake, Bastian. 2017. “Looking into Pandora’s Box: The Content of Sci-Hub and Its Usage.” *F1000Research* 6 (April): 541. https://doi.org/10.12688/f1000research.11366.1.

Himmelstein, D. S., A. R. Romero, J. G. Levernier, T. A. Munro, S. R. McLaughlin, B. Tzovaras Greshake, and C. S. Greene. 2018. “Sci-Hub Provides Access to Nearly All Scholarly Literature.” *ELife* 7. https://doi.org/10.7554/eLife.32822.

Holdren, John P. 2013. “Increasing Access to the Results of Federally Funded Scientific Research.” Washington D. C.: Office of Science and Technology Policy, The White House.

Hungarian Academy of Sciences Electronic Information National Programme Service. 2018. “EISZ - Hungarian Consortium Terminates Negotiations with Elsevier.” 2018. http://eisz.mtak.hu/index.php/en/283-hungarian-consortium-terminates-negotiations-with-elsevier.html.

Karaganis, Joe, ed. 2018. *Shadow Libraries - Access to Knowledge in Global Higher Education*. Cambridge, MA: MIT Press.

Krikorian, Gaëlle, and Amy Kapczynski, eds. 2010. *Access to Knowledge in the Age of Intellectual Property*. Zone Books.

Kwon, Diana. 2018. “Universities in Germany and Sweden Lose Access to Elsevier Journals.” TheScientist. 2018. https://www.the-scientist.com/news-opinion/universities-in-germany-and-sweden-lose-access-to-elsevier-journals--64522.

Liaw, Andy. 2020. *RandomForest v4.6-14 - Breiman and Cutler’s Random Forests for Classification and Regression*.

Machin-Mastromatteo, Juan D, Alejandro Uribe-Tirado, and Maria E Romero-Ortiz. 2016. “Piracy of Scientific Papers in Latin America: An Analysis of Sci-Hub Usage Data.” *Information Development* 32 (5): 1806–14.

Mars, Marcell, and Tomislav Medak. 2015. “The System of a Takedown - Control and De-Commodification in the Circuits of Academic Publishing.”

Max Planck Society. 2018. “Max Planck Society Discontinues Agreement with Elsevier; Stands Firm with Projekt DEAL Negotiations - MPDL.” 2018. https://www.mpdl.mpg.de/en/505.

Molnar, Christoph. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Molnar, Christoph, Bernd Bischl, and Giuseppe Casalicchio. 2018. “Iml: An R Package for Interpretable Machine Learning.” *Journal of Open Source Software* 3 (26): 786. https://doi.org/10.21105/joss.00786.

Muller, Floriane Sophie, and Pablo Iriarte. 2017. “Measuring the Impact of Piracy and Open Access on the Academic Library Services.” In *15th Interlending and Document Supply Conference (ILDS) , Paris, 04-06 October, 2017*. Paris.

Poort, Joost, João Quintais, Martin A. van der Ende, Anastasia Yagafarova, and Mathijs Hageraats. 2018. “Global Online Piracy Study.” Amsterdam: Institute for Information Law, University of Amsterdam.

Reimers, Imke. 2014. “The Effect of Piracy Protection in Book Publishing.” NBER.

Suber, Peter. 2013. *Open Access*. Vol. 1. Cambridge, MA: The MIT Press. https://doi.org/10.1109/ACCESS.2012.2226094.

Swartz, Aaron. 2008. “Guerilla Open Access Manifesto.” Aaron Swartz. 2008. https://archive.org/stream/GuerillaOpenAccessManifesto/Goamjuly2008\_djvu.txt.

Taylor, Alasdair. 2006. “Publishing and Electronic Piracy.” *Learned Publishing* 19 (3): 168–74.

Timus, Natalia, and Babutsidze Zakaria. 2016. “Pirating European Studies.” *Journal of Contemporary European Research* 12 (3).

Travis, J. 2016. “In Survey, Most Give Thumbs-up to Pirated Papers.” Science. 2016. http://www.sciencemag.org/news/2016/05/survey-most-give-thumbspirated-papers.

UC Office of the President. 2019. “UC Terminates Subscriptions with World’s Largest Scientific Publisher in Push for Open Access to Publicly Funded Research.” University of California. 2019. https://www.universityofcalifornia.edu/press-room/uc-terminates-subscriptions-worlds-largest-scientific-publisher-push-open-access-publicly.

Unit. 2019. “Norske forskningsinstitusjoner har besluttet å ikke forlenge avtale med forlaget Elsevier.” 2019. https://www.unit.no/aktuelt/norske-forskningsinstitusjoner-har-besluttet-ikke-forlenge-avtale-med-forlaget-elsevier.

1. The research received funding from the H2020 Research grant “[OPENing UP new methods, indicators and tools for peer review, dissemination of research results, and impact measurement](http://openup-h2020.eu/)”, and was carried out on the Dutch national e-infrastructure with the support of [SURF Cooperative](https://www.surf.nl/en/homepage). [↑](#footnote-ref-1)
2. At the time of writing in May, 2019, 2.363.587 records are contained in the database [↑](#footnote-ref-2)
3. Such as file hashes [↑](#footnote-ref-3)
4. We do not discuss here recent developments in the business strategies of academic publishers, which increasingly rely on capturing, processing and reselling data on, and tools of academic knowledge production process from citation metrics, bibliographic tools, altmetric services, article discovery tools, scientific social networks, as well as performance metrics, valorization metrics, and impact measurement. [↑](#footnote-ref-4)
5. https://www.maxmind.com/en/geoip-demo [↑](#footnote-ref-5)
6. We removed repeated requests from the same IP address to the same title if they took place within 24 hours. [↑](#footnote-ref-6)
7. The source of macro data on population (SP.POP.TOTL), GDP (per capita, PPP, current international $ - NY.GDP.PCAP.PP.CD) , Fixed broadband subscriptions (IT.NET.BBND), Literacy rate (adult total , % of people ages 15 and above - SE.ADT.LITR.ZS), Research and development expenditure (% of GDP - GB.XPD.RSDV.GD.ZS), and School enrollment, tertiary (% gross - SE.TER.ENRR) is the <https://data.worldbank.org/> database. The source of Government expenditure on tertiary education per student in constant 2014 PPP US$ 2015 is the OECD’s Education at a Glance database, <http://www.oecd.org/education/education-at-a-glance/>, while the source of H indices is the Scimago Journal & Country Rank dataset <https://www.scimagojr.com/countryrank.php> [↑](#footnote-ref-7)
8. Data source: <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519> [↑](#footnote-ref-8)
9. The problem with territorially aggregated data is that such data is mainly available on a country level. That is a serious limitation, because many variables that can affect the environment of pirated book downloads is not available for many countries, which renders a more nuanced application impossible. Another problem is that country size is very heterogeneous. The United States is technically one country of 52 states, and almost the size of the European Union with 28 countries. Smaller countries are less heterogeneous and less complex from an aggregation point of view. For a good understanding of environmental factors, we would like to see a high level of heterogeneity across territories and a low level of heterogeneity within territories, and the EU’s NUTS statistical system allows this enhanced view. [↑](#footnote-ref-9)
10. We did not exclusively work with NUTS2 level. Obvious exceptions were the very small member states, like Malta, or Estonia, where the country size does not allow a distinction between NUTS0 (national) NUTS1 (larger regional) and NUTS2 (regional) data. In these cases, we used only technically country-level data, which is not different from the NUTS1 or NUTS2 level data, as the territory of these small states is not further divided on these levels. The other exceptions were those European comparative survey-based newer statistical products, where the way the survey sample was constructed, the regional break-up in larger countries, like in Germany or in the Great Britain part of the United Kingdom, were only available in NUTS1 levels. See the detailed data descriptions in the supplementary materials. [↑](#footnote-ref-10)
11. Eurostat statistics were programmatically downloaded with the rOpenGov package (Lahti et al. 2019): (Eurostat 2019b, 2019a), (Eurostat 2019c). [This is only testing, we need to find a good way to display 52 variable citations.] [↑](#footnote-ref-11)
12. Eurobarometer is a series of public opinion surveys conducted on behalf of the European Commission on a wide range of issues. <https://www.gesis.org/eurobarometer-data-service/home/> [↑](#footnote-ref-12)
13. We ran several models which included the R&D expenditure instead of / next to the researcher headcount, but there was no model in which the R&D expenditure was significant. [↑](#footnote-ref-13)
14. The TOR browser allows users to preserve their online privacy by routing their online traffic through a number of intermediary computers to a random exit point on the internet. This has the implication that traffic from such TOR exit nodes, though associated with a particular IP address and geographic location usually originate from elsewhere. [↑](#footnote-ref-14)
15. We used the randomForest R package (Liaw 2020) for this purpose. First, we established the optimal parameters for starting the algorithm with the tuneRF function. We used all predictors to build a forest of regression trees. We used the Interpretable Machine Learning method and package (Molnar, Bischl, and Casalicchio 2018) to interpret the importance of each feature (see particularly Chapter 5.5. of (Molnar 2019)). [↑](#footnote-ref-15)
16. The other models are in the supplementary materials. [↑](#footnote-ref-16)
17. See supplementary materials for the details. [↑](#footnote-ref-17)
18. The limited library access is a weighted sum of the responses that gave the answer “Limited or poor quality of this activity in the place where you live” to the question: “Please tell me why you haven’t visited a public library?” [↑](#footnote-ref-18)