Estimating the Real with the Virtual: How does the Accuracy Relate to the Internet Penetration?

Ignacio Castro
Queen Mary University of London
London, UK

Sergey Gorinsky
IMDEA Networks Institute
Madrid, Spain

***Abstract*—Direct measurements of the real world are typically expensive and time-consuming. Recent studies have revealed the promising quick alternative where real-world characteristics are indirectly estimated based on data readily available in the Internet. Such methods for indirect virtual-world estimation of real-world properties rely on the implicit hypothesis that the estimation accuracy depends on how extensively the Internet penetrates into the real world. Our paper makes the first step towards validating this hypothesis. We adopt a simple statistical model to quantify the relationship between an index of consumer prices and online searches for respective categories of goods and services across 19 countries. We also show how the strength of this relationship depends on Internet-penetration and other socioeconomic variables.**

I. Introduction

Collection of reliable data about a real-world phenomenon can take significant amounts of time and resources. While useful for medical practice, social policy, and other important applications, data collection via health surveys, opinion polls, market studies, and government censuses generally puts humans in the loop, causing delays and increasing costs. Is there an alternative to such costly slow methods?

The Internet has grown into a global virtual infrastructure that produces and stores vast information about the real world. Banking, shopping, entertainment, and social networking are increasingly done online, and data related to these virtual-world activities are recorded in various repositories. Can the virtual-world activity information be useful for measuring the real world?

A number of studies have successfully utilized online-activity data to indirectly estimate properties of real-world phenomena. While Google became the dominant search engine in many countries, Google Trends is an online service that reports popularity of various terms in Google searches [1]. Scientists employed Google Trends to track disease outbreaks [2], software engineering practices [3], stock market fluctuations [4], unemployment benefit claims, travel, car, and home sales [5], private consumption [6], as well as correlation between GDP (gross domestic product) and future user interests [7]. Also, data from Twitter, a social-networking and microblogging service [8], were used to estimate stock market dynamics [9], [10].

The success of the indirect estimation approach depends on how closely the real-world phenomenon and virtual-world activity are correlated. For example, the intensity of Google queries about a particular travel destination is shown to be closely related with how popular the travel destination is in reality [5]. Hence, the real-world travel can be accurately estimated based on the respective virtual-world search data from Google Trends.

In estimating a real-world phenomenon with online-activity data, there is an implicit hypothesis that the estimation accuracy is related to how extensively the Internet penetrates into the real world [11]–[13]. The intuition behind this previously not studied hypothesis is that for the real-world phenomenon and virtual-world activity to be strongly correlated, the online activity should not only be of a sufficiently relevant kind but also involve a sufficiently large number of users. Conversely, when the Internet penetration is shallow, the online-activity data are unlikely to reflect the real-world phenomenon accurately. Meanwhile, the Internet penetration is greatly uneven around the world, e.g., the population percentage using the Internet in 2011 varied from 0.3% in Sierra Leone to 95% in Iceland [14].

In this paper, we make the first step towards validating the hypothesis that the accuracy of estimating the real with the virtual depends on the Internet penetration. In our statistical analysis, the estimated real-world phenomenon is consumer prices, as represented by their harmonized indices from Eurostat [15], and the estimating online-activity data are results from Google Trends for respective categories of consumed goods and services. The study has both geographical and historical dimensions: the data cover 19 countries from 2005 to 2011. We also show how the estimation accuracy depends on Internet-penetration measures, such as the broadband-penetration rate and Internet-access frequency [13], and other socioeconomic variables. The main contributions of this paper are as follows:

- We are the first to study the hypothesis that the accuracy of real-world phenomena with online-activity data depends on the Internet penetration into the real world.
- Our statistical analysis covers 19 countries over 7 years and considers Internet-penetration indicators and auxiliary socioeconomic variables.
- We show that, while not particularly strong, the relationship between the estimation accuracy and Internet penetration is tangible, with a potential to strengthen the relationship by accounting for additional factors.

978-1-4673-9622-6/16/$31.00 ©2016 IEEE

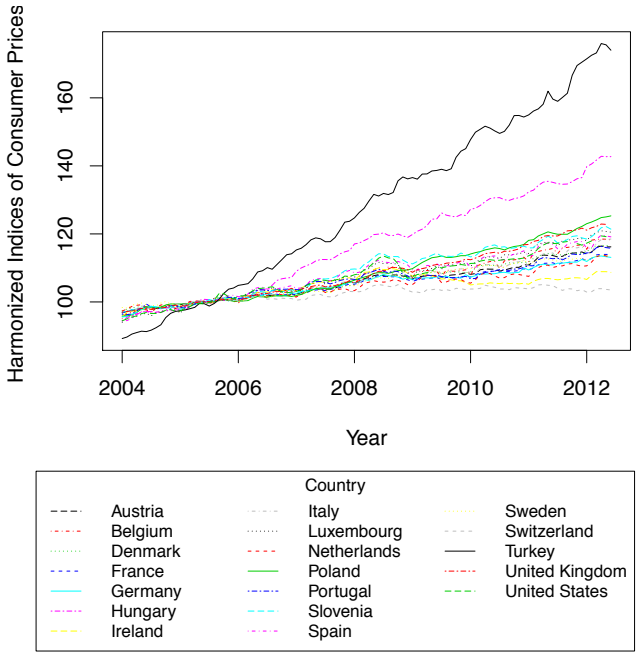


Fig. 1: Dynamics of consumer-price indices, the real-world phenomenon data in our study

The rest of this paper has the following structure. Section II presents the data used in our study. Section III reports the evaluation methodology and results. Section IV reviews related work. Section V outlines directions for future work. Finally, section VI concludes the paper with a summary.

II. Data

Our investigation into the relationship between the estimation accuracy and Internet penetration involves real-world, online-activity, Internet-penetration, and auxiliary socioeconomic data. These 4 sets of data are presented in Sections II-A, II-B, II-C, and II-D respectively.

A. Real-world phenomenon data

The major constraint in choosing a real-world phenomenon for our study is availability of periodic high-quality data for the phenomenon across multiple countries. After considering IMF (International Monetary Fund), WB (World Bank), UN (United Nations), OECD (Organisation for Economic Co-operation and Development), Eurostat, which is the statistical office of the EU (European Union) [15], and other sources of cross-country data, we select consumer prices as the studied real-world phenomenon, with the respective data provided by Eurostat's HICPs (Harmonized Indices of Consumer Prices) [16]. HICPs are comparable economic indicators of price stability for consumer goods and services acquired by households. To harmonize the prices paid by households for the goods and services, the index computations use weights that account for different expenditures in different categories. Eurostat reports HICPs at the monthly granularity for the EU countries, Turkey, and USA. The legend box in Figure 1 lists the 19 countries included in our analysis. The figure depicts the dynamics of HICPs over the period from 2004 to 2012. While the price indices exhibit inflationary patterns in all considered countries, the inflation is the steepest in Hungary and Turkey.

Category number	Category name
1	Arts or Entertainment
2	Autos or Vehicles
3	Beauty or Fitness
4	Books or Literature
5	Business or Industrial
6	Computers or Electronics
7	Finance
8	Food or Drink
9	Games
10	Health
11	Hobbies or Leisure
12	Home or Garden
13	Internet or Telecom
14	Jobs or Education
15	Law or Government
16	News
17	Online Communities
18	People or Society
19	Pets or Animals
20	Real Estate
21	Reference
22	Science
23	Shopping
24	Sports
25	Travel

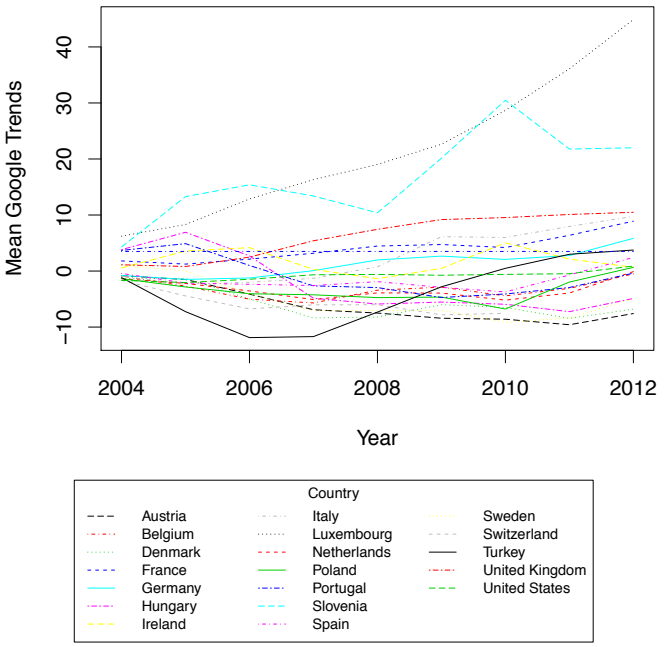
TABLE I: Utilized categories of Google Trends results, the online-activity data in our analysis

B. Online-activity data

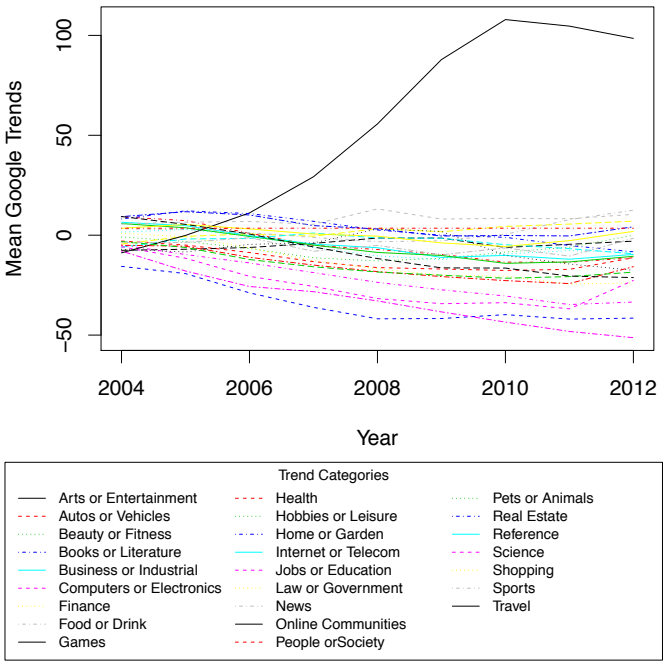
Because the willingness of consumers to buy goods and services is correlated with Google queries for the products [6], we employ cross-country results from Google Trends [1] as the online-activity data in our study. This Google service reports on relative popularity of terms in Google searches. Negative values for a term in Google Trends mean its popularity decreases in comparison to other searches. For some countries, Google Trends divides its weekly data into 25 categories related to consumer goods and services. Table I shows these 25 categories. While Google Trends reports both actual data and forecasts, our analysis relies only on the actual data for the 25 categories. We aggregate the online-activity data from their original weekly granularity to the monthly resolution, i.e., to the same temporal granularity as our real-world phenomenon data. Figure 2 depicts the dynamics of the used online-activity data.

C. Internet-penetration data

The notion of Internet penetration is broad and quantified via a number of diverse metrics [11]–[13]. Public sources



(a) for all considered countries



(b) for Austria in specific categories

Fig. 2: Dynamics of the the online-activity data

for Internet-penetration data are also diverse. WB offers data on the number of Internet users for most of the world countries from early 1990s [17]. Internet-usage information is also available from the ITU (International Telecommunication Union) World Telecommunication database [14]. We choose Eurostat [15] because it provides a rich data set with many different variables covering a wide range of Internet-penetration aspects. Additionally, the Eurostat data are already harmonized and thus lend themselves readily for our analytical purposes. Table II presents the Internet-penetration metrics from the used Eurostat data set. The metric names, e.g., **isoc_tc_broad**, are kept the same as originally assigned by Eurostat. Additional details about the Eurostat Information Society statistics and survey questionnaires are respectively available in [18] and [19].

D. Auxiliary socioeconomic data

To examine whether other socioeconomic factors affect the relationship between the consumer prices and online-activity data, we consider such additional properties as the population age, educational level, and wealth, which are commonly used to understand Internet penetration. To obtain the auxiliary data, we again benefit from Eurostat. Table III defines the 4 used auxiliary Eurostat metrics: **demo_pjanind**, **demo_pjan**, **educ_ipart_s**, and **nama_aux_gph**. Our last auxiliary variable **search_engine** comes from a different source, a public web-analysis service [20]. This variable measures the percentage of Internet users relying on Google as their search engine. The time series for the metric start in 2008 and are depicted in figure 3. The **search_engine** variable tracks the representativeness of Google as the main search engine. The metric is relevant to our study because our online-activity data come from Google Trends. Note that Google is neither the only search engine nor dominant in every country, e.g., Yandex [21] dominates Google in Russia [22]. On the other hand, except for the United States and Slovenia, at least 90% of Internet users in the considered countries use Google for web search.

III. Evaluation

A. Estimating the real with the virtual

Time-series models are common for predicting real-world phenomena with online-activity data [5], [9]. While linear regressions are prone to the endogeneity problem that undermines their predictive power, time-series analyses are excellent for forecasting and do not explicitly deal with underlying causes. In our econometric model, we use VAR (Vector Autoregression) to capture the linear interdependencies among the time series for the considered countries and predict HICPs y_t at time t with exogenous Google Trends inputs x_{t-i} estimated with OLS (Ordinary Least Squares):

$$y_t = C + \sum_{i=1}^{n} A_i x_{t-i} + \varepsilon_t \qquad (1)$$

where $t = 1, \ldots, T$ with T representing the length of the time series, i is a lag of the autoregressive order, n denotes the maximum lag, C and A_i are the intercept vector and parameter matrix respectively, and ε_t refers to the vector of disturbances. Because we aim at result comparability, our selection of the VAR method over more complex techniques, such as ARIMA (AutoRegressive Integrated Moving Average), is for reducing potential bias in the model choice according to the Akaike information criterion (AIC).

Variable number	Variable name	Variable description
1	**isoc_tc_broad**	Broadband penetration rate: the number of dedicated connections with a capacity equal to or higher than 144 Kbps per 100 inhabitants
2	**isoc_ci_in_h**	Number of households with Internet access
3	**isoc_ci_ifp_iu**	Percentage of individuals using the Internet in the last 12 months
4	**isoc_bde15cua**	Percentage of individuals using the Internet at least once a week
5	**isoc_bde15cbc**	Online banking indicator: the percentage of individuals using online banking

TABLE II: Internet-penetration metrics

Variable number	Variable name	Variable description
1	**demo_pjan**	Total population
2	**demo_pjanind**	Median age of the population
3	**educ_ipart_s**	Students at ISCED (International Standard Classification of Education) level 3 (upper and secondary education) as a percentage of the total number of students
4	**nama_aux_gph**	GDP per capita
5	**search_engine**	Percentage of Internet users relying on Google as their search engine

TABLE III: Auxiliary socioeconomic variables

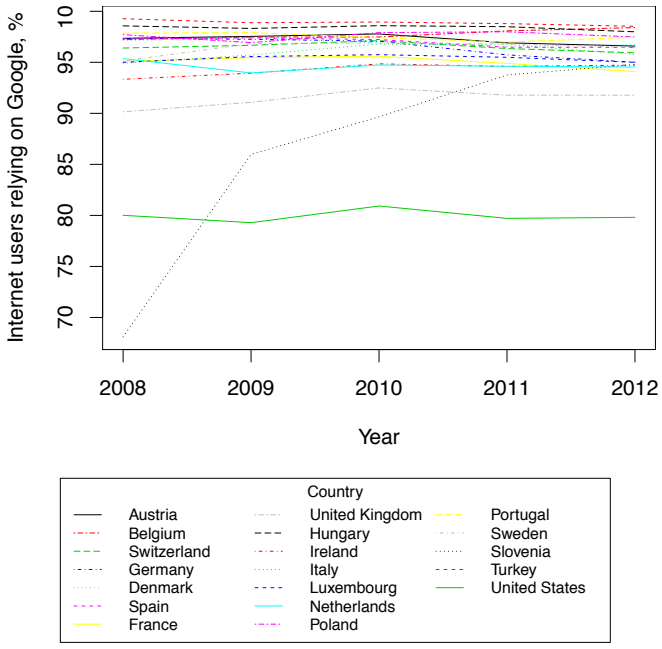


Fig. 3: Percentage of users relying on Google as their search engine (auxiliary variable **search_engine**)

B. Estimation accuracy

To measure the estimation accuracy, we look for a scale-free metric that enables comparisons of the results across different countries. Fulfilling this goal, we select an accuracy metric based on MASE (Mean Absolute Scaled Error) that uses a random walk over the T-element time series for the fitted data to scale error e_t at time t [23]. Specifically, we define the estimation accuracy as a negation of MASE so that a more accurate estimate corresponds to a larger value of the metric:

$$\textbf{accuracy} = -\text{mean}\left(\frac{|e_t|}{\frac{1}{T-1}\sum_{i=2}^{T}|y_i - y_{i-1}|}\right). \quad (2)$$

The **accuracy** metric is well-defined and finite (except in the irrelevant case where all observations are equal). It takes non-positive values with 0 denoting the completely accurate estimate. When the prediction is more accurate than the benchmark, the metric value exceeds -1.

While other metrics of estimation accuracy exist, they can give infinite or undefined values in commonly occurring situations [23]. More importantly, these metrics do not effectively support our need to compare results across different models. Whereas we also consider an alternative MASE-based metric, this paper does not report the respective results because they are qualitatively similar.

C. Methodology

In our estimation of the real with the virtual via equation 1, the dimensions of the used C and A_i (i.e., intercept vector and parameter matrix) are respectively 19 and 19×25. To keep the computations simple, we limit the maximum lag of the autoregressive order to $n = 1$. For each considered country, we estimate a model for the standardized data from 2005 to 2011, i.e., all years with complete data availability. We conduct such estimation with the R-CRAN statistical software [24] and, in particular, its EvalEst [25] and DSE [26] packages. Unlike the prior work on estimating real-world phenomena with online-activity data, our approach does not focus on maximizing the estimation accuracy. Instead, we emphasize result comparability and do all estimations in the same statistical model that includes all the online-activity variables.

To study how the estimation accuracy is correlated with Internet-penetration and auxiliary variables, we again rely on

	accuracy	**isoc_ci_in_h**	**isoc_tc_broad**	**isoc_ci_ifp_iu**	**isoc_bde15cbc**	**isoc_bde15cua**
(Intercept)	-338.91 ***	-39.30 **	-97.58 ***	-33.55 *	-41.81 **	-102.29 ***
	(87.58)	(18.97)	(19.01)	(19.70)	(20.95)	(34.18)
nama_aux_gph	$1 \cdot 10^{-3}$ ***	$4 \cdot 10^{-7}$ ***	$4 \cdot 10^{-3}$ ***	$4 \cdot 10^{-7}$ ***	$4 \cdot 10^{-7}$ ***	$4 \cdot 10^{-8}$ ***
	$(4 \cdot 10^{-3})$	$(4 \cdot 10^{-1})$	$(4 \cdot 10^{-1})$	$(4 \cdot 10^{-1})$	$(4 \cdot 10^{-1})$	$(4 \cdot 10^{-1})$
educ_ipart_s	0.26	-0.06	0.14 **	-0.02	-0.08	0.09
	(0.26)	(0.09)	(0.05)	(0.09)	(0.10)	(0.11)
demo_pjanind	7.26 ***	2.08 ***	2.57 ***	2.04 ***	1.46 ***	3.46 ***
	(1.07)	(0.42)	(0.43)	(0.41)	(0.46)	(0.78)
demo_pjan	$8.68 \cdot 10^{-8}$	$-1.86 \cdot 10^{-8}$	$-5.82 \cdot 10^{-8}$ *	$-2.28 \cdot 10^{-8}$	$-2.96 \cdot 10^{-8}$	$-1.15 \cdot 10^{-7}$ **
	$(1.3 \cdot 10^{-7})$	$(4.77 \cdot 10^{-8})$	$(2.99 \cdot 10^{-8})$	$(4.80 \cdot 10^{-8})$	$(5.40 \cdot 10^{-8})$	$(5.56 \cdot 10^{-8})$
search_engine	-0.11					
	(0.73)					
R^2	0.64	0.53	0.35	0.51	0.46	0.43
Adjusted R^2	0.5711	0.5123	0.3361	0.4921	0.4466	0.4181
Number of observations	54	114	109	112	111	108

TABLE IV: Explanation of the accuracy and Internet-penetration variables with the auxiliary socioeconomic factors

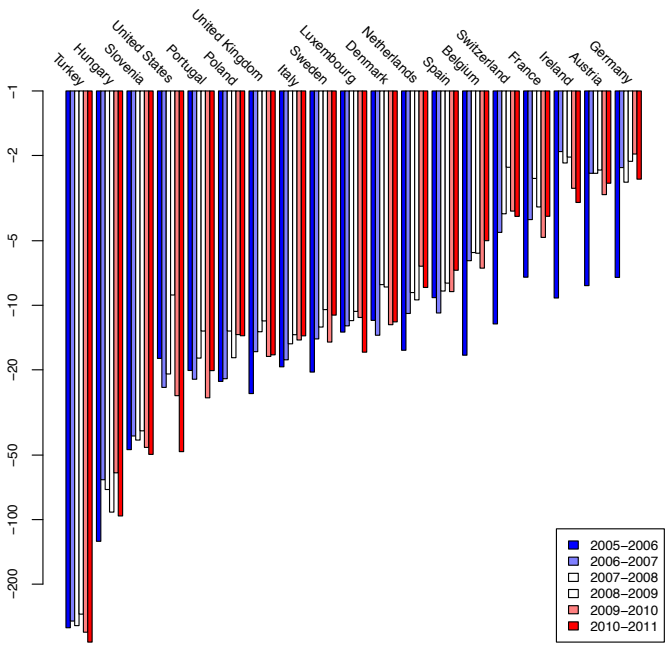


Fig. 4: Estimation accuracy for the considered countries

regression techniques. The techniques use pooled OLS and data from all considered countries and years:

$$y_{jt} = \gamma + \beta x_{jt} + u_{jt} \tag{3}$$

where j refers to a country, t denotes time, y is the explained variable, x is the vector of explanatory variables, γ is an intercept, β represents the vector of estimated coefficients, and u_{jt} is the error term.

We capture the correlation between the compared variables by reporting the estimated coefficients and R^2, the coefficient of determination [27]. Also, we use an additional metric of **adjusted** R^2 that adjusts R^2 for the number of explanatory terms in the model. Both R^2 and **adjusted** R^2 take values between 0 and 1, with 1 indicating a good fit of the regression and data. The metrics can be interpreted as the percentage of the explained variable that the model explains. Additionally, we employ p-values to judge whether the coefficients of the explanatory variables are significant [27]. The p-test checks the hypothesis that each coefficient differs from 0. We consider p-values below 0.01, which corresponds to the significance levels between 90% and 99%. These are the levels where a variable is typically considered statistically significant.

D. Results

First, we evaluate the estimation accuracy for each of the 19 studied countries over the 2005-2011 period. Figure 4 plots **accuracy** for every year within this period. All the observed values are substantially below -1. Whereas the estimation accuracy is generally low, the accuracy varies dramatically from country to country. The accuracy is the lowest in Turkey and the highest in Germany. In general, West European countries tend to have higher **accuracy** than countries in the European East and South. The accuracy in the United States is lower as well. In the temporal dimension, the accuracy tends to increase in countries where it is higher. Interestingly, the tendency is opposite in countries where the accuracy is lower: **accuracy** tends to decrease there.

After quantifying the estimation accuracy, we examine how **accuracy** relates to our five Internet-penetration variables: **isoc_tc_broad**, **isoc_ci_in_h**, **isoc_ci_ifp_iu**, **isoc_bde15cua**, and **isoc_bde15cbc**. The study is done by regressing the accuracy and Internet-penetration variables with the auxiliary indicators of the per-capita GDP (**nama_aux_gph**), education level (**educ_ipart_s**), median population age (**demo_pjanind**), total population (**demo_pjan**). The regression for **accuracy** also considers the Google usage (**search_engine**).

Table IV sums up the numerical results of these 6 regressions, with the table rows and columns representing respec-

	accuracy
(Intercept)	-574.83 *** (141.87)
isoc_ci_in_h	20.41 *** (6.09)
isoc_ci_in_h^2	-0.25 ** (0.10)
isoc_ci_in_h^3	$1.1 \cdot 10^{-3}$ * $(4 \cdot 10^{-6})$
search_engine	0.23 (0.79)
R^2	0.55
Adjusted R^2	0.51
Number of observations	67

(a)

	accuracy
(Intercept)	-1021.03 *** (239.94)
isoc_ci_ifp_iu	41.46 *** (10.64)
isoc_ci_ifp_iu^2	-0.57 *** (0.16)
isoc_ci_ifp_iu^3	0.26 *** $(4 \cdot 10^{-8})$
search_engine	0.08 (0.81)
R^2	0.52
Adjusted R^2	0.48
Number of observations	67

(b)

	accuracy
(Intercept)	-227.73 *** (77.20)
isoc_bde15cbc	12.85 *** (1.83)
isoc_bde15cbc^2	-0.27 *** (0.04)
isoc_bde15cbc^3	0.18 *** $(4 \cdot 10^{-8})$
search_engine	0.31 (0.73)
R^2	0.62
Adjusted R^2	0.58
Number of observations	66

(c)

	accuracy
(Intercept)	-63.47 (65.05)
isoc_tc_broad	0.98 ** (0.39)
search_engine	0.21 (0.67)
R^2	0.09
Adjusted R^2	0.09
Number of observations	64

(d)

	accuracy
(Intercept)	-56.79 (67.31)
isoc_bde15cua	0.31 (0.20)
search_engine	0.18 (0.69)
R^2	0.04
Adjusted R^2	0.04
Number of observations	64

(e)

TABLE V: Explanation of the estimation accuracy with the Internet-penetration variables

tively the explanatory and explained variables. The numerical values in a table cell denote the coefficient value and (in parentheses) standard error. We identify the statistical significance of the variable with the star notation: 1, 2, and 3 stars denote respectively the confidence level of 90%, 95% and 99%. In general, the results show that **accuracy** and Internet-penetration variables behave similarly with respect to the auxiliary indicators: the socioeconomic variables that are significant for the Internet-penetration measures are also significant for the estimation accuracy; the coefficients have the same sign and similar values; the overall explanatory power of the models is comparable.

Zooming in on specific explanatory variables, we see that the per-capita GDP and median population age are significant in all cases and appear with the positive sign, i.e., the estimation accuracy and Internet penetration are higher for richer and older populations. Education is only significant for the broadband-penetration metric and has the expected positive sign. The total population offers little explanatory power in the considered models. This auxiliary variable is significant in explaining only the broadband-penetration rate and appears with a very small coefficient. R^2 ranges from 0.39 for the broadband-penetration metric to 0.64 for **accuracy**. The values for R^2 and **adjusted** R^2 remain consistently close to each other, implying that the models do not include unnecessary variables.

To directly explore the relationship between the estimation accuracy and Internet penetration, we conduct 5 pooled OLS regressions for **accuracy** where different subsets of the Internet-penetration measures serve as the explanatory variables. Table V shows results for these 5 regressions and supports a general conclusion that while the estimation accuracy and Internet-penetration variables are certainly related, the relationship is not particularly strong.

In regard to specific Internet-penetration metrics, the number of households with Internet access (**isoc_ci_in_h**) and Internet usage in the last 12 months (**isoc_ci_ifp_iu**) have the largest correlation with **accuracy**. Using a third-order polynomial to exploit their explanatory power, we are able to explain more than a half of the data, i.e., $R^2 \geqslant 0.5$. As expected, the signs are positive for the first-order polynomial and negative for the quadratic form of the variable. This reveals the typical decreasing returns for the explanatory power of the variable. The broadband-penetration rate (**isoc_tc_broad**) and Internet usage once a week (**isoc_bde15cua**) are weakly related with the estimation accuracy.

Due to the recent emergence of new Internet access alternatives such as mobile access technologies, the poor predictive

	accuracy
(Intercept)	20371.17 *** (4618.75)
isoc_ci_ifp_iu	-276.39 ** (131.17)
search_engine	-268.64 *** (63.25)
isoc_bde15cbc	-192.33 ** (72.43)
isoc_ci_in_h	-117.15 ** (44.92)
isoc_ci_ifp_iu^2	1.60 *** (0.46)
search_engine^2	0.56 *** (0.18)
isoc_bde15cbc^2	0.35 (0.34)
isoc_ci_ifp_iu×**search_engine**	3.06 ** (1.34)
search_engine×**isoc_bde15cbc**	1.93 ** (0.73)
search_engine×**isoc_ci_in_h**	1.15 ** (0.45)
isoc_ci_ifp_iu^2 × **search_engine**^{2}	$-4 \cdot 10^{-2}$ *** ($4.82 \cdot 10^{-6}$)
search_engine^2 × **isoc_bde15cbc**^{2}	$-4 \cdot 10^{-1}$ ($3.72 \cdot 10^{-5}$)
isoc_ci_ifp_iu×**isoc_bde15cbc**	0.07 (0.15)
isoc_bde15cbc×**isoc_ci_in_h**	0.18 *** (0.05)
R^2	0.84
Adjusted R^2	0.79
Number of observations	66

TABLE VI: Explanation of the estimation accuracy with the expanded set of variables

power of the broadband-penetration rate is not surprising. In contrast, the Internet usage in the last 12 months is a metric that covers all Internet access methods, and its correlation with the estimation accuracy is much stronger. The observation that **accuracy** is much stronger correlated with the total Internet user population (**isoc_ci_ifp_iu**) than with the user group that accesses the Internet frequently (**isoc_bde15cua**) makes sense because the price indices of real-world consumption are closer related to the overall number of consumers/users, rather than to the size of a particular group.

Taking a closer look at those variables that performed better in the previously examined regressions of **accuracy**. Table VI presents the main results for a pooled regression with such explanatory variables. Surprisingly, the individual Internet-penetration metrics appear as strongly significant but with the negative sign. The positive sign of the quadratic form suggests that this effect has a decreasing tendency. Also, the negative sign is reversed when we look at interaction variables, i.e., products of the variables. The sign reversal also occurs with the quadratic interaction variables, denoting again that the contribution of the variable decreases as the variable increases. A general message of table VI is that while Internet penetration alone does not have an especially strong correlation with the estimation accuracy, accounting for additional factors can strengthen the relationship. For example, whereas **isoc_ci_ifp_iu** and **isoc_bde15cbc** (online banking) affect the estimation accuracy negatively when considered independently, the product of **isoc_ci_ifp_iu** and **isoc_bde15cbc** has a significant positive effect, and the model accounting for their interaction is able to explain 75% of the estimation accuracy.

IV. Related Work

In 2005, [28] unveiled the potential of online-activity data for estimation of real-world phenomena. Assuming that queries in web search engines reflect "needs, wants and interests", [28] used search-engine keywords to estimate the number of unemployed workers in the United States. Subsequently, data from search engines and social-media services, such as Twitter, became common means for estimating a wide variety of real-world properties, e.g., stock market fluctuations [9], [10] and geographical presence of Mexican drug cartels [29]. After [5] championed Google Trends [1] as a general tool for estimation of socioeconomic statistics, researchers have used Google Trends to estimate disease outbreaks [2], software engineering practices [3], stock market fluctuations [4], unemployment benefit claims, travel, car, and home sales [5], private consumption [6], as well as correlation between GDP and future interests of Internet users [7].

Despite the success stories, online-activity data are not necessarily representative of real-world phenomena in general. Also, a specific usage of the Internet can be only weakly related to the estimated phenomenon. For instance, movie box-office estimations with Twitter data were found to be of a poor quality unless combined with data from movie-rating websites such as IMDb [30]. In the context of our study, Internet users do not necessarily represent well the general consumers characterized by HICPs. In comparison with Twitter, Google Trends is more likely to capture a larger and more general group of users and thus partially avoids this problem [30].

While our estimation-accuracy metric is based on MASE, there exist other measures of estimation accuracy, e.g., RMSE (Root Mean Squared Error) or MAE (Mean Absolute Error) [5]. These alternatives suffer from well-known drawbacks, e.g., RMSE and MAE are not scale-free measures, leading to potential inaccuracies in comparisons [23].

V. Future work

A natural future extension for our work is to examine more countries over longer time intervals. While Google Trends provides weekly 25-category data for the considered countries, such category-specific data are unavailable for some countries, especially for developing ones. This is unfortunate because considering both developed and developing countries is likely

to yield better insights into the relationship between the estimation accuracy and Internet penetration.

Another direction for future work is to utilize more sophisticated statistical techniques, including advanced machine learning algorithms, to increase the estimation accuracy, especially at the level of individual countries. For example, it is worth to explore Monte Carlo methods with different combinations of Google Trend categories and lag structures. Note though that the overall goal for our work is not to develop the most accurate estimation of a real-world phenomenon with online-activity data but to understand how strongly the estimation accuracy and Internet penetration are correlated.

VI. Conclusions

A number of recent studies have used online-activity data, e.g., from Google searches or Twitter social networking, to estimate real-world phenomena such as consumption of goods and services. In this paper, we tested the implicit hypothesis of the previous studies that the accuracy of such estimations is correlated with Internet penetration. Our statistical evaluation used consumer prices (Eurostat's HICPs) as the estimated real-world phenomenon, and Google Trends for the corresponding online-activity data. The data covered 19 countries and 25 categories that captured the willingness of consumers to buy goods and services.

In the geographical dimension, we showed that West European countries tended to have a higher estimation accuracy than countries in the European East and South. In the historical dimension, while the estimation accuracy tended to increase in countries where it was higher, the tendency was opposite in countries where the estimation accuracy was lower. Our results revealed that the estimation accuracy and Internet-penetration variables behaved similarly with respect to the auxiliary socioeconomic indicators, e.g., the estimation accuracy and Internet penetration were higher for richer and older populations. Overall, we showed that the correlation between the estimation accuracy and Internet penetration was tangible but not especially strong, with a potential to strengthen the relationship by accounting for additional factors. The insufficiently strong correlation undermined one of the original ambitions of our work: if the correlation were strong, the estimation accuracy could be used as a new metric of Internet penetration. Future extensions of the work with different choices of data and statistical methods might bring this intent closer to fruition.

VII. Acknowledgments

We are grateful to Juan Camilo Cardona for the code to collect the Google Trends data. We also thank Suzy Moat and Tobias Lechtenfeld for their help in positioning our work. This research was financially supported in part by the European Commission (FP7-ICT 288021, EINS), Regional Government of Madrid (S2013/ICE-2894, Cloud4BigData), and Spanish Ministry of Science and Innovation (TEC2014-55713-R, HyperAdapt).

References

[1] Google Trends. http://www.google.es/trends.
[2] H. A. Carneiro and E. Mylonakis, "Google Trends: A Web-based Tool for Real-time Surveillance of Disease Outbreaks," *Clinical Infectious Diseases*, vol. 49, no. 10, pp. 1557–1564, November 2009.
[3] J. Rech, "Discovering Trends in Software Engineering with Google Trend," *ACM SIGSOFT Software Engineering Notes*, vol. 32, no. 2, pp. 1–2, March 2007.
[4] T. Preis, D. Reith, and H. E. Stanley, "Complex Dynamics of Our Economic Life on Different Scales: Insights from Search Engine Query Data," *Philosophical Transactions of the Royal Society*, vol. 368, no. 1933, pp. 5707–5719, November 2010.
[5] H. Choi and H. Varian, "Predicting the Present with Google Trends," *Economic Record*, vol. 88, no. s1, pp. 2–9, June 2012.
[6] S. Vosen and T. Schmidt, "Forecasting Private Consumption: Survey Based Indicators vs. Google Trends," *Journal of Forecasting*, vol. 30, no. 6, pp. 565–578, September 2011.
[7] T. Preis, H. S. Moat, H. E. Stanley, and S. R. Bishop, "Quantifying the Advantage of Looking Forward," *Scientific Reports*, vol. 2, April 2012.
[8] Twitter, https://twitter.com.
[9] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, March 2011.
[10] J. Bollen, A. Pepe, and H. Mao, "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-economic Phenomena," *arXiv:0911.1583*, pp. 1–10, November 2009.
[11] M. D. Chinn and R. W. Fairlie, "The Determinants of the Global Digital Divide: a Cross-country Analysis of Computer and Internet Penetration," *Oxford Economic Papers*, vol. 59, no. 1, pp. 16–44, January 2007.
[12] S. Kiiski and M. Pohjola, "Cross-country Diffusion of the Internet," *Information Economics and Policy*, vol. 14, no. 2, pp. 297–310, June 2002.
[13] I. Peña-López, "Towards a Comprehensive Model of the Digital Economy," in *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, December 2010.
[14] International Telecommunications Union (United Nations), "Statistics," https://www.itu.int/ITU-D/ict/statistics.
[15] Eurostat (European Union), http:/ec.europa.eu/eurostat.
[16] ——, "Compendium of HICP Reference Documents," http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-13-017.
[17] World Bank, "Data," http://data.worldbank.org/.
[18] Eurostat (European Union), "Information Society Statistics - Households and Individuals," http://ec.europa.eu/eurostat/statistics-explained/index.php/Information_society_statistics_-_households_and_individuals.
[19] ——, https://circabc.europa.eu/webdav/CircaBC/ESTAT/emisannexes/Library/data_-_database/theme_3_-_popul/isoc/householdsindividuals.
[20] StatCounter Global Stats, http://gs.statcounter.com.
[21] Yandex, https://www.yandex.ru.
[22] Alexa, "Top Sites in Russia," http://www.alexa.com/topsites/countries/RU.
[23] R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, November 2006.
[24] R Core Team, "The R Project for Statistical Computing," https://www.r-project.org.
[25] Gilbert, P., "EvalEst Guide," http://cran.r-project.org/web/packages/EvalEst/vignettes/Guide.pdf.
[26] ——, "Brief User's Guide: Dynamic Systems Estimation (DSE)," https://cran.r-project.org/web/packages/dse/vignettes/Guide.pdf.
[27] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.
[28] M. Ettredge, J. Gerdes, and G. Karuga, "Using Web-based Search Data to Predict Macroeconomic Statistics," *Communications of the ACM*, vol. 48, no. 11, pp. 87–92, November 2005.
[29] M. Coscia and V. Rios, "How and Where do Criminals Operate? Using Google to Track Mexican Drug Trafficking Organizations," http://www.gov.harvard.edu/files/videos/CosciaRios_GoogleForCriminals.pdf, October 2012.
[30] F. M. F. Wong, S. Sen, and M. Chiang, "Why Watching Movie Tweets won't Tell the Whole Story?" in *Proceedings of the 2012 ACM Workshop on Online Social Networks*, August 2012.