Utrecht University

School of Economics

The Pearls and Perils of Google Trends

A Housing Market Application

Joep Steegmans

U.S.E. Working Paper Series nr: 19-11

Utrecht University School of Economics

Utrecht University School of Economics (U.S.E.) is part of the faculty of Law, Economics and Governance at Utrecht University. The U.S.E. Research Institute focuses on high quality research in economics and business, with special attention to a multidisciplinary approach. In the working papers series the U.S.E. Research Institute publishes preliminary results of ongoing research for early dissemination, to enhance discussion with the academic community and with society at large.

The research findings reported in this paper are the result of the independent research of the author(s) and do not necessarily reflect the position of U.S.E. or Utrecht University in general.

U.S.E. Research Institute
Kriekenpitplein 21-22, 3584 EC Utrecht, The Netherlands Tel: +31 30 253 9800, e-mail: use.ri@uu.nl www.uu.nl/use/research

Utrecht University

School of Economics

U.S.E. Research Institute
Working Paper Series 19-11

The Pearls and Perils of Google Trends: A Housing Market Application

Joep Steegmans
Utrecht School of Economics
Utrecht University

August 2019

Abstract
This study aims to provide insights into the correct usage of Google search data, which are available through Google Trends. The main focus is on the effects of sampling error in these data as these are ignored by most scholars using Google Trends. To demonstrate the effect a housing market application is used; that is, the relationship between online search activity for mortgages and real housing market activity is investigated. A simple time series model, based on Van Veldhuizen, Vogt, and Voogt (2016), is estimated that explains house transactions using Google search data for mortgages. The results show that the effects of sampling errors are substantial. It is also stressed that in this particular application of Google Trends data 'predetermined' transactions, house sales where the purchase contracts have been signed but where the conveyance hasn't occurred yet, should be excluded as they lead to an overestimation of the effects of mortgage searches. All in all, the application of Google Trends data in economic applications remains promising. However, far more attention should be given to the limitations of these data.

Keywords: Google Trends; internet search; housing market

JEL classification: D83, L86, R21

Comments welcomed to: J.W.A.M.Steegmans@uu.nl

1. Introduction

In recent years various scholars have looked into the possibilities of linking online search behavior to real market activity, particularly through the use of Google search queries. The best known examples are focusing on unemployment (Askitas and Zimmermann 2009), consumption (Vosen and Schmidt 2011), and stock markets (e.g. Preis, Moat, and Stanley 2013). The housing market and Google search queries have been linked by Choi and Varian (2009), Wu and Brynjolfsson (2015), and Van Veldhuizen, Vogt, and Voogt (2016).[1] This current paper demonstrates some important limitations in forecasting real housing market activity with online search data. To do so, the model provided in Van Veldhuizen, Vogt, and Voogt (2016) will be used.

The paper provides insights into the correct usage of Google Trends (which provides the Google search data) as it does recognize the potential of online search data in housing market applications. The paper will focus on two aspects: sampling error and causality. I will demonstrate that the estimations by Van Veldhuizen, Vogt, and Voogt (2016) overestimate the importance of Google Trends data in the prediction of housing market transactions. Nevertheless, I also show that excluding the 'predetermined' transactions, in which causality is likely to run the other way around, does not necessarily render the Google Trends data useless. I will demonstrate that including Google Trends data can lead to improvement compared to a benchmark model where online search activity is ignored.

The paper contributes to the literature on applying big data sources in economic research. Particularly, the paper adds insights to housing market applications of Google Trends data. More than any other housing market study it stresses the limitations in using these data. The rest of this brief article is organized as follows. Section 2 provides a short summary of Van Veldhuizen, Vogt, and Voogt (2016). Section 3 looks into sampling errors, which are inherent to Google Trends data. Section 4 focusses on causal sequentiality by differentiating between house purchases/sales and conveyances. Section 5 presents the estimation results in which sampling error and causal sequentiality are taken into account. Section 6 summarizes and concludes.

2. Van Veldhuizen, Vogt & Voogt (2016)

Van Veldhuizen, Vogt, and Voogt (2016) relate Google searches for mortgages to housing transactions. They estimate a simple linear time series model where monthly transactions on the macro level are explained by aggregate Google searches, and up to 11 of its lags.

The monthly transaction data cover the period from January 2004 until October 2015. The transaction data are publicly available at Statistics Netherlands (CBS StatLine 2015). The search data, queries for the Dutch word for mortgage (i.e. *hypotheek*), are obtained from Google Trends (2004-2018) and cover the same period. The search data are obtained at a weekly level and are aggregated by the authors into monthly data in the eventual analysis. Google Trends data provide an index of the search query: the week in which the relative usage of the query, compared to the total number of queries, is highest is set to 100, while all other periods are expressed as a ratio of this maximum.

[1] It should be noted that only the Working Paper version of the study by Choi and Varian contained an analysis of homes sales. In the published version, i.e. Choi and Varian (2012), home sales have been dropped in its entirety.

The starting point of Van Veldhuizen, Vogt, and Voogt (2016) is a time series model that excludes online search activity; in this benchmark model transaction numbers are simply corrected for seasonality and time trends.

$$y_t = \alpha + \gamma \boldsymbol{T} + \epsilon_t \tag{1}$$

where y_t indicates the standardized number of monthly transactions (i.e. conveyances) and $\boldsymbol{T}$ includes the set of both year and month dummies.

The benchmark model is extended to include search activity as an additional predictor.

$$y_t = \alpha + \beta \boldsymbol{X}_t + \gamma \boldsymbol{T} + \epsilon_t \tag{2}$$

where the matrix $\boldsymbol{X}_t$ includes online search activity of either a month or a year. More precisely, $\boldsymbol{X}_t$ exists of the standardized Google Trends index of the mortgage queries and up to 11 of its lags. The preferred specification of Van Veldhuizen, Vogt, and Voogt (2016) is current search activity plus search activity in the 11 preceding months.

Van Veldhuizen, Vogt, and Voogt (2016) conclude that (i) Google searches for mortgages in current and previous months are "highly significantly" positively associated with housing transactions, (ii) mortgage searches six and nine months prior are significantly positively associated with housing transactions, and (iii) including mortgage searches increases the explanatory power of the simple prediction model of housing transactions by 4 percentage points (p. 1321).

For this paper I will initially use the exact same data as Van Veldhuizen, Vogt, and Voogt (2016). Apart from that, I have extended the data set with 100 additional Google Trends indexes for mortgage search activity in the Netherlands, which were downloaded on 100 consecutive days (see Section 3). One important difference is that I have downloaded the Google Trends data as a monthly series; that is, I do not have to transform the weekly series into a monthly one.[2]

3. Sampling error

The first issue that I would like to address is the sampling error in Google Trends data. The sampling error occurs because Google uses only "a percentage of searches" to compile the index (Google 2018). Information on the sample size, however, is not disclosed by Google. A complication herein is that new samples can be taken only once per day as data is cached by Google on a daily basis only (Stephens-Davidowitz and Varian 2015). In other words, Google Trends uses a different sample to generate the index on every single date (e.g. McLaren and Shanbhogue 2011).

Thus far scholars have given little attention to the sampling error in Google Trends data. This might partly be contributed to Stephens-Davidowitz and Varian (2015) who have stated that they "do not expect that [...] researchers will need more than a single sample" (p. 13). This current paper stresses that such a general claim should not be made if only because Trends data depend on geography and time span. Nevertheless, most scholars ignore the sampling error altogether.

[2] At the present time Google Trends provides only monthly data for periods longer than five years, while before the data was provided at a weekly level of observation. Van Veldhuizen, Vogt, and Voogt (2016) use the the first day of the week to determine in what month the weekly observation is included.

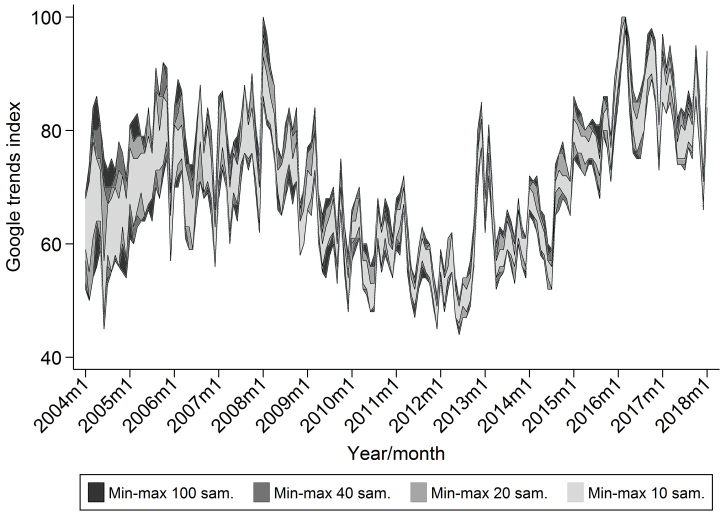

Figure 1. Min-max range for up to 100 Google Trends samples, queries for "hypotheek" the Netherlands.

Figure 1 illustrates the sampling error of the mortgage query by plotting the minimum and maximum values for each month for 100 samples, taken on 100 consecutive days. In the figure different sample sizes have been superimposed, demonstrating that the min-max range increases when the number of samples increases. Between January 2004 and October 2015, the period under investigation, the min-max range varies between 6 and 30 index points or, in relative terms, between 8.9 percent and 66.7 percent. Figure 1 indicates that the sampling error can be very substantial.

Van Veldhuizen, Vogt, and Voogt (2016) are among the studies that rely on a single sample to draw inference. The risks of relying on a single sample are demonstrated by re-estimating the model specification of Van Veldhuizen, Vogt, and Voogt (2016) with indexes based on 100 separate Google Trends samples.

Table 1 shows the signs of the estimated coefficients of Equation (2) from Van Veldhuizen, Vogt, and Voogt (2016), based on their single sample, and the estimated signs for the 100 Google Trends indexes that have been newly collected. The table indicates that at least part of the findings of Van Veldhuizen, Vogt, and Voogt (2016) can be contributed to the specific sample that was used for their Google Trends index. More particularly, it has not been possible to confirm the finding that the sixth and ninth lag of mortgage searches are significantly positively related to housing transactions. Table 1 shows a positive coefficient for the sixth lag in only 6 of the 100 estimations and in 82 of the 100 estimations for the ninth lag. Furthermore, it turns out that finding positive coefficients for both is even rarer: in only 3 of the 100 estimations the coefficients of both the sixth *and* the ninth lag are positive, indicating that the estimated coefficients are not independent of each other.

4. Financing conditions and causality

The second issue that needs addressing is related to the transaction variable from the Cadastre records that is used by Van Veldhuizen, Vogt, and Voogt (2016). They use the date of conveyance (completion), the date at which ownership is transferred from one party to the other. Importantly, the date at which the final offer is accepted (which

Table 1. Comparison of Van Veldhuizen et al. (2016) specification with 100 additional samples.

	Single sample VVV2016	Repeated sampling Google Trends			
	Sign	Positive coef.	Zero	Negative coef.	Samples
Google searches t	+	95	5	0	100
Google searches t-1	+	97	3	0	100
Google searches t-2	0	12	88	0	100
Google searches t-3	0	13	87	0	100
Google searches t-4	0	0	99	1	100
Google searches t-5	0	0	100	0	100
Google searches t-6	+	6	94	0	100
Google searches t-7	0	0	100	0	100
Google searches t-8	0	0	100	0	100
Google searches t-9	+	82	18	0	100
Google searches t-10	0	0	100	0	100
Google searches t-11	0	58	42	0	100

Notes: The dependent variable is the standardized number of transactions (i.e. conveyances). A 10 percent significance level is used.

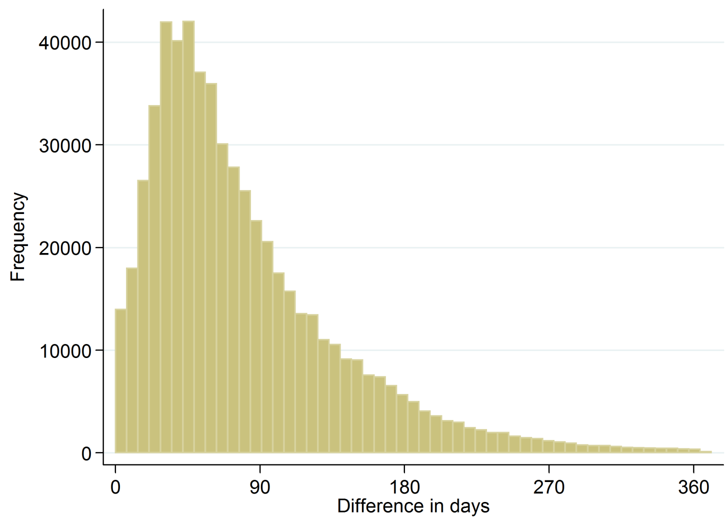

Figure 2. Distribution of time between sales agreement and conveyance for family homes.

is legally binding in the Netherlands) and the date at which the purchase contracts are signed precede the date of conveyance.

I have compared transaction dates of the Cadastre with transaction dates of the Dutch Association of Realtors to illustrate the aforementioned issue. Figure 2 depicts the time between the conveyance date and the closing of the listing by the realtor, which generally is the date the purchase contract is signed, for a subsample of 584,923 family homes between January 2004 and August 2013 for which I am able to observe both. The figure illustrates that the purchase date does not coincide with the conveyance date. On average the difference is about three months, as is the median.

The differences in the realtor and Cadastre transaction dates are explained by both necessity and preference: the buyer needs to arrange financing and parties will have personal preferences regarding the moment to move in or out. Agreements in the Netherlands are almost exclusively reached under specific conditions: most importantly, purchase contracts include financing conditions. The clause makes the contract void if the buying party is not able to arrange mortgage financing. The Dutch association of owner-occupiers (VEH) states that in the Netherlands a period of 6 to 8 weeks

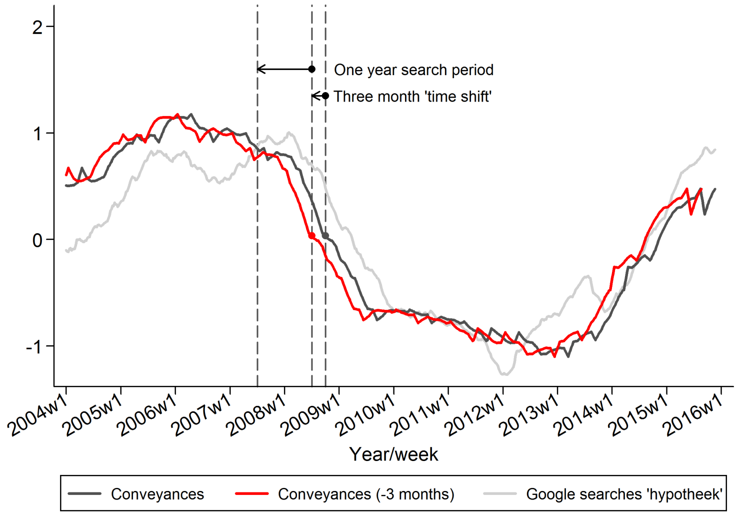

Figure 3. Times series of house transactions (conveyance) and search trend.

is common to arrange a mortgage (Vereniging Eigen Huis s.a.).

The financing condition specified in the purchase contract is of particular interest as Van Veldhuizen, Vogt, and Voogt (2016) try to predict house transactions based on internet search behaviour. While I agree that aggregate online search on mortgage information might have predictive power for future house transactions, causality runs predominantly the other way after the purchase agreement has been signed. It is the purchase of a house that causes the buyer to search for mortgages. Hence it makes no sense to include search data from after the moment the purchase has been agreed upon to predict transactions.

Figure 3 illustrates the issue by making use of the data used in Van Veldhuizen, Vogt, and Voogt (2016).[3] The figure emphasizes that houses where ownership was conveyed in week 40 of 2008 (chosen for illustrative purposes only) on average transacted three months earlier (week 27 of 2008). A one year search period would thus, on average, run from fourteen months till three months before the conveyance. I thus suggest to exclude online searches for mortgages within the three-month window prior to the conveyance when 'predicting' Cadastre transactions.

5. Results

As explained above the specification of main interest will exclude online search activity in period t and the first two lags. The specification with one year search thus includes the third until the fourteenth lag of the Google Trends index. Table 2 shows the summary of the estimation results of Equation (2) for both the sample of Van Veldhuizen, Vogt, and Voogt (2016) and the additional 100 Google Trends indexes that were collected. Excluding the 'predetermined' transactions suggests only limited evidence of individual coefficients being significant. Focusing on the 100 newly collected samples (columns 2-4), the third lag is significant in 80 of the 100 estimations and the ninth

[3]For illustrative purposes I present the moving averages of the times series (ranging from minus to plus 36 weeks), as do Van Veldhuizen, Vogt, and Voogt (2016). In the analyses the non-smoothed, more volatile data are used.

Table 2. Comparison of Van Veldhuizen et al. (2016) with 100 additional samples, excl. predetermined transactions.

	Single sample VVV2016	Repeated sampling Google Trends			
	Sign	Positive coef.	Zero	Negative coef.	Samples
Google searches t-3	0	80	20	0	100
Google searches t-4	0	0	97	3	100
Google searches t-5	+	0	100	0	100
Google searches t-6	–	6	94	0	100
Google searches t-7	0	0	100	0	100
Google searches t-8	0	2	98	0	100
Google searches t-9	0	61	39	0	100
Google searches t-10	0	0	100	0	100
Google searches t-11	0	1	99	0	100
Google searches t-12	0	3	97	0	100
Google searches t-13	0	0	98	2	100
Google searches t-14	–	0	82	18	100

Notes: The dependent variable is the standardized number of transactions (i.e. conveyances). A 10 percent significance level is used.

lag is significant in 61 of the estimations. However, the importance of Table 2 is not in determining predictability of housing transactions based on mortgage searches, it is – once again – to demonstrate the effects of sampling error. The table demonstrates that using a particular sample can have major consequences in the findings. This also follows from the estimates in the first column of Table 2, based on the VVV2016 sample: the coefficients seem to suggest that more online mortgage searches could decrease the number of transactions.[4]

In order to draw conclusions on predictability of housing transactions the average of the 100 Google Trends indexes is used. Table 3 shows the estimation results of the benchmark model, the one-month search extension, and the twelve-month search extension. The third column shows the results for a one-month search period. The third lag is barely significant (p-value is 0.0742), while the adjusted R-squared increases with 0.3 percentage points compared to the benchmark (1.4 p.p. in the original study).

The last two columns in Table 3 show that the third lag (p-value is 0.0270) and the ninth lag (p-value is 0.0731) are significant or on the verge of being significant. Testing joint significance of the lags provides only limited evidence of the lags being relevant: the p-value of the F-test is 0.100. The adjusted R-squared of the one year search model increases with 1.1 percentage points compared to the benchmark (3.9 p.p. in the original study). The table shows that excluding the predetermined transactions, i.e. increases in search activity after the purchase contract has been signed, leads to much smaller effects than found by Van Veldhuizen, Vogt, and Voogt (2016).

Comparing Table 2 with Table 3 does provide an illustration of the effects of sampling error. Comparing the p-values of the significant lags (Table 3) with the probability of finding a positive coefficient in one of the sampled indexes (Table 2) suggests that for the individual estimations the coefficients are biased towards zero; the is, the measurement error in the individual Google Trends samples leads to attenuation bias. All in all, there is only little evidence that Google Trends data is useful in predicting housing transactions.

[4]The estimated coefficients for the VVV2016 sample, including the benchmark and one-month search specification, can be found in Table A1 in the appendix.

Table 3. Estimated results excluding predetermined transactions (averaged Google Trends index).

	(1) Benchmark		(2) One month search (lag 3)		(3) One year search (lags 3-14)	
Google searches t-3			0.1662*	(0.0742)	0.2241**	(0.0270)
Google searches t-4					-0.1316	(0.2265)
Google searches t-5					-0.0518	(0.6284)
Google searches t-6					0.0838	(0.4296)
Google searches t-7					0.0410	(0.7012)
Google searches t-8					0.0363	(0.7335)
Google searches t-9					0.1918*	(0.0731)
Google searches t-10					-0.0657	(0.5272)
Google searches t-11					0.0766	(0.4549)
Google searches t-12					0.1207	(0.2378)
Google searches t-13					-0.0721	(0.4857)
Google searches t-14					-0.1591	(0.1230)
Constant	0.2965	(0.1185)	0.2112	(0.2746)	0.2688	(0.2360)
Year dummies	Yes		Yes		Yes	
Month dummies	Yes		Yes		Yes	
Observations	128		128		128	
Adj. R-squared	0.834		0.837		0.845	
P-value joint sign.			0.074		0.100	

Notes: The dependent variable is the standardized number of transactions (i.e. conveyances). P-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

6. Conclusion

This paper focuses on the correct usage of Google Trends data in housing market applications. To do so the paper looks into the relationship between online search activity for mortgages and real housing market activity. Some of the pitfalls in Google Trends applications are demonstrated by re-estimating the model of Van Veldhuizen, Vogt, and Voogt (2016), who study the relation between online mortgage searches and housing transactions. This current paper argues that the strong correlation between transaction numbers and online search found by the aforementioned authors is due to a combination of sampling error and a misspecified causal relationship.

The sampling error in Google Trends data is studied by collecting an additional 100 samples of the Google Trends index for the Dutch word for mortgage and re-estimating the model of Van Veldhuizen, Vogt, and Voogt (2016). The estimation results for the 100 additional samples lead in only 3 of the 100 samples to the same findings, i.e. a significant sixth and ninth lag for search activity. Hence I argue that their findings are based on the peculiarity of their one sample.

I also argue that Van Veldhuizen, Vogt, and Voogt (2016) should have distinguished between the signing of the purchase contract and the conveyance, which follows on average three months later. The financing condition in purchase contracts allows households to arrange mortgage financing after the purchase contract has been signed. The signing of the contract thus leads to increased interest in mortgages. It makes no sense to include search activity after a purchase has been agreed upon to predict transactions based on the conveyance date. Search activity in the three months prior to the conveyance should therefore not be included when predicting house transactions. A 12-month search period should thus include the third until the fourteenth lag of search.

After excluding the 'predetermined' transactions there is little evidence that online search leads to higher transaction numbers. The preferred model, which makes use of the mean of the 100 newly collected Google Trends indexes, is not entirely conclusive. In the 1-month search specification the third lag of mortgage search is barely significant (p-value is 0.0742). In the 12-month search specification the third and the ninth lag seem significant (p-values of 0.0270 and 0.0731, respectively). Still, the p-value of the joint significance of all twelve lags seems not entirely conclusive (p-value is 0.100). At best one can conclude that both the 1-month search specification and the 12-month search specification have a slightly higher explanatory power than the benchmark model where search is not included as a predictor (0.3 and 1.1 percentage points, respectively). All in all, I conclude that the relationship between online search activity and transaction numbers is much weaker than suggested by Van Veldhuizen, Vogt, and Voogt (2016).

This study has stressed the limitations in using Google search data (i.e. Google Trends). I still recognize the enormous potential of online search data and believe that alternative time series models could do a better job in predicting house transactions than the simple model that was applied here. Nevertheless, the focus was not on finding the best prediction model of house transactions based on searches for mortgages. The main lesson is that due to drawbacks in both the construction of Google Trends data and the sampling method these data should only be used very cautiously.

Acknowledgements

This research has been made possible through financial support of the Ministry of the Interior and Kingdom Relations (the Netherlands). I would like to thank Benedikt Vogt for providing me with the original data and code. Furthermore, I would like to thank Wolter Hassink and Marc Schramm for providing me with useful comments.

References

Askitas, Nikolaos, and Klaus F Zimmermann. 2009. "Google econometrics and unemployment forecasting." *Applied Economics Quarterly* 55 (2): 107–120.

CBS StatLine. 2015. "Bestaande koopwoningen; verkoopprijzen prijsindex 2010 = 100 [Existing houses; sales price index 2010 = 100]. Data file." Centraal Bureau voor de Statistiek (Statistics Netherlands), The Hague. http://statline.cbs.nl/Statweb/.

Choi, Hyunyoung, and Hal Varian. 2009. *Predicting the present with Google Trends.* Working paper (april 10, 2009). Google Inc. Available at SSRN: https://ssrn.com/abstract=1659302.

Choi, Hyunyoung, and Hal Varian. 2012. "Predicting the present with Google Trends." *Economic Record* 88 (Special Issue): 2–9.

Google. 2018. "Where Trends data comes from." Google Trends Help Center. https://support.google.com/trends/answer/4355213?hl=en (last visited on 03/05/2018).

Google Trends. 2004-2018. "Zoekterm hypotheek; Interesse in de loop der tijd [Search term 'hypotheek'; interest over time]; 2004-2018 [Data file]." Google Trends. https://trends.google.nl/trends/.

McLaren, Nick, and Rachana Shanbhogue. 2011. "Using internet search data as economic indicators." *Bank of England. Quarterly Bulletin* 51 (2): 134.

Preis, Tobias, Helen Susannah Moat, and H Eugene Stanley. 2013. "Quantifying trading behavior in financial markets using Google Trends." *Scientific reports* 3: srep01684.

Stephens-Davidowitz, Seth, and Hal Varian. 2015. *A Hands-on Guide to Google Data.* Unpublished manuscript.

Van Veldhuizen, Sander, Benedikt Vogt, and Bart Voogt. 2016. "Internet searches and transactions on the Dutch housing market.” *Applied Economics Letters* 23 (18): 1321–1324.

Vereniging Eigen Huis. s.a. *Verkocht onder voorbehoud [Sold subject to finance].* Webpage. Dutch association of owner-occupiers. www.eigenhuis.nl/huis-verkopen/stappenplan-huis-verkopen/huis-verkocht/verkocht-onder-voorbehoud (last visited on 11/01/2018).

Vosen, Simeon, and Torsten Schmidt. 2011. “Forecasting private consumption: survey-based indicators vs. Google trends.” *Journal of Forecasting* 30 (6): 565–578.

Wu, Lynn, and Erik Brynjolfsson. 2015. “The future of prediction: How Google searches foreshadow housing prices and sales.” In *Economic analysis of the digital economy*, 89–118. University of Chicago Press.

Appendix A.

Table A1. Re-estimated results excluding predetermined transactions (original data: Jan 2004–Oct 2015).

	(1) Benchmark		(2) One month search (lag 3)		(3) One year search (lags 3-14)	
Google searches t-3			0.1374	(0.1067)	0.1269	(0.1620)
Google searches t-4					0.0058	(0.9510)
Google searches t-5					-0.1688*	(0.0780)
Google searches t-6					0.1829**	(0.0421)
Google searches t-7					-0.0514	(0.5664)
Google searches t-8					0.0632	(0.4729)
Google searches t-9					0.1347	(0.1262)
Google searches t-10					0.0118	(0.8931)
Google searches t-11					-0.0108	(0.9008)
Google searches t-12					0.0969	(0.2755)
Google searches t-13					0.0102	(0.9090)
Google searches t-14					-0.1893**	(0.0413)
Constant	0.2965	(0.1185)	0.1501	(0.4708)	0.1532	(0.5064)
Year dummies	Yes		Yes		Yes	
Month dummies	Yes		Yes		Yes	
Observations	128		128		128	
Adj. R-squared	0.834		0.836		0.845	
P-value joint sign.			0.107		0.097	

Notes: The dependent variable is the standardized number of transactions (i.e. conveyances). P-values in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.