zef
Center for
Development Research
University of Bonn

ZEF-Discussion Papers on
Development Policy No. 271

Regine Weber and Lukas Kornher

Can one improve now-casts of crop prices in Africa? Google can.

Bonn, February 2019

Weber, Regine, Kornher, Lukas, Can one improve now-casts of crop prices in Africa? Google can., ZEF – Discussion Papers on Development Policy No. 271, Center for Development Research, Bonn, February 2019, pp. 36.

ISSN: 1436-9931

Published by:
Zentrum für Entwicklungsforschung (ZEF)
Center for Development Research
Genscherallee 3
D – 53113 Bonn
Germany
Phone: +49-228-73-1861
Fax: +49-228-73-1869
E-Mail: zef@uni-bonn.de
www.zef.de

The authors:
Regine Weber, Center for Development Research (ZEF), University of Bonn.
Contact: rweber@uni-bonn.de
Lukas Kornher, Center for Development Research (ZEF), University of Bonn.
Contact: lkornher@uni-bonn.de

Acknowledgements

The authors thank Prof. Dr. Joachim von Braun, Dr. Heike Baumüller, Dr. Bernardina Algieri and Dr. Elías Cisneros for their valuable comments and suggestions. Kieran Munelly provided excellent research assistance. The authors gratefully acknowledge the financial support from the Federal Ministry for Economic Cooperation and Development (BMZ) under the Center for Development Research (ZEF) Project “Measures to reduce price volatility for improved food security”.

Abstract

With increasing Internet user rates across Africa, there is considerable interest in exploring new, online data sources. Particularly, search engine metadata, i.e. data representing the contemporaneous online-interest in a specific topic, has gained considerable interest, due to its potential to extract a near real-time online signal about the current interest of a society. The objective of this study is to analyze whether search engine metadata in the form of Google Search Query (GSQ) data can be used to improve now-casts of maize prices in nine African countries, these are Ethiopia, Kenya, Malawi, Mozambique, Rwanda, Tanzania and Uganda, Zambia and Zimbabwe. We formulate as benchmark an auto-regressive model for each country, which we subsequently augment by two specifications based on contemporary GSQ data. We test the models in in-sample, and in a pseudo out-of-sample, one-step-ahead now-casting environment and compare their forecasting errors. The GSQ specifications improve the now-casting fit in 8 out 9 countries and reduce the now-casting error between 3% and 23%. The largest improvement of maize price now-casts is achieved for Malawi, Kenya, Zambia and Tanzania, with improvements larger than 14%.

Keywords: Food Security, Maize Prices, Search Query Volume, Internet, Early Warning Systems

JEL classification: Q18, Q11, O33

1. Introduction

With the emergence of the Internet, new, online data sources have become available, as people produce digital traces when using the Internet. This online metadata, which is usually aggregated over a vast body of Internet users, contains a signal derived from a larger number of people than usually covered by surveys. In that regard, particularly search engine metadata, i.e. data representing the contemporaneous online-interest in a specific topic, or more specifically, what people currently search for as they navigate the Internet, has gained considerable interest. Tapping into this kind of information holds the potential to extract a near real-time online signal about the current interest of a society.

Across many African countries, Internet-adoption rates have started to increase significantly and more than doubled in many countries over the past decade. Average Internet-user rates currently range at around 24% of the population (International Telecommunication Unit 2018). This development coincides with a persistent risk to food security, driven by *inter alia* recurrent droughts, extreme weather events and conflicts. Early warning systems and situation monitoring play a crucial role in decision making processes and facilitate preventive actions and early interventions. Early warning and situation monitoring requires fast, disaggregated and reliable information, to produce timely forecasts and potential warnings. In many African countries, however, high-frequency information is difficult to obtain and official statistics are published with a considerable time lag, at lower frequency and quality (Kalkuhl, von Braun, and Torero 2016). Hence, decision makers face the challenge of having to make decisions in scenarios where information is lacking (Carrière-Swallow and Labbé 2013). Given these factors, particularly in the context of developing countries, extracting a near real-time online signal about the contemporaneous interest of a society could help identifying upcoming crises and has the potential to contribute to and improve current models and decision making processes. Therefore, there is considerable interest also in Africa, to explore the prospect of online, high-frequency information for now- and short-term forecasting models, i.e. models that predict the present or the very near future (Bańbura et al. 2013).

In the realm of search engine metadata, Google search query (GSQ) data is of particular interest, due to Google's dominance in the search engine market and its search engine metadata being published free of charge. GSQ data reflects the search volume of a specific keyword entered into the Google search engine at a certain location and point in time, hence, representing the contemporaneous online interest in a specific topic. GSQ data holds promising potential for the now-casting and inter-period forecasting of a variety of indicators, since Google releases its query data on a weekly basis and, hence, earlier than standard reports and data used for crises forecasting. The use of GSQ data has found wide applications during the last decade: from understanding the spread of epidemics (Ginsberg et al. 2009; Lazer et al. 2014), to political attitudes (Stephens-Davidowitz 2013; Marthews and Tucker 2014) and human behavior (Stephens-Davidowitz 2017), as well as in the field of economics, to now-casting and forecasting private consumption (Vosen and Schmidt 2011), inflation expectations (Guzmán 2011), stock market volatility (Hamid and Heiden 2015), developments on financial markets (Preis, Moat, and Stanley 2013), exchange rates (Bulut 2017), and unemployment rates (Askitas and Zimmermann 2015; Suhoy 2009). These studies, however, share one aspect: the use of GSQ data in the context of industrialized countries, where high Internet-adoption rates prevail. Two notable exemptions are Carrière-Swallow and Labbé (2013), who use GSQ data to now-cast automobile sales in Chile as well as Seabold and Coppola (2015), who now-cast consumer price indices and staple food prices in Costa Rica, El Salvador and Honduras.

To date, we are unaware of any attempt that explores the link between food price developments, as a proxy indicator for food security, and online-signals in the form of search query data in Africa, i.e. in an environment with relatively low Internet-adoption rates. The objective of this study is to address this research gap and to answer the research question whether GSQ data can be used to now-cast maize prices in a selection of African countries. This study does not aim to seek a substitute for price data, it rather seeks to investigate whether models including GSQ data can serve as a proxy for price developments. Our study focuses on nine African countries, which we selected based on a data driven approach. These are Ethiopia, Kenya, Malawi, Mozambique, Rwanda, Tanzania and Uganda, Zambia and Zimbabwe.

2. Studies using GSQ Data

Various disciplines have explored GSQ data to predict the present and near future. In the field of epidemics, Ginsberg et al. (2009) use a non-public data set of GSQ data to monitor flu trends in the US, while Lazer et al. (2014) develop an improved flu map based on public GSQ data. GSQ data has further been used to explore people's attitudes towards sensitive topics that are either not covered by surveys or that are usually prone to be over- or under-reported. Stephens-Davidowitz (2013) *inter alia* develops a GSQ measure for racial animosity in the US to analyze the percentage points of Barack Obama’s forgone turnout in the 2008 presidential election. He finds that Obama lost 8 % due to racial animosity, a larger estimate compared to traditional survey estimates of racial bias. Marthews and Tucker (2014) use GSQ data to analyze the attitude towards Internet privacy of the US's top 40 trading partners before and after the PRISM revelations, i.e. information leaks about the large-scale surveillance program of the US National Security Agency. Their findings show that post PRISM, search engine behavior changed in relation to sensitive queries, such as health queries and that this effect on search engine behavior is more pronounced in countries that are usually considered US allies.

In the field of economics, GSQ data has been used for the intra-period forecasting of economic indicators and consumer sentiments. Choi and Varian (2012) show that the inclusion of GSQ data in simple auto-regressive models of automobile sales, unemployment claims, travel-destination planning and consumer confidence, improves the model fit and that models with Google data outperform models without Google data by 5 to 20%. The forecasting capacity of Google Trends with regards to unemployment rates has further been analyzed for Germany (Askitas and Zimmermann 2009) and Israel (Suhoy 2009). Vosen and Schmidt (2011) show that the forecasting performance of private consumption in the US can be improved by including an index based on Google search queries. They find that the Google index outperforms standard survey based indicators, like the University of Michigan Consumer Sentiment Index and the Conference Board Consumer Confidence Index, in both in- and out-of-sample consumption forecasts.

With regards to studies on financial markets, Guzmán (2011) analyzes the predictive power of various standard measures of inflation expectations in the US as well as the Google search volume for the keyword *inflation*, with focus on differences in data frequency. She finds the GSQ indicator to have the lowest out-of-sample forecasting error. Preis, Moat, and Stanley (2013) analyze the relationship between the Google search volume and financial markets in the US. They find that the Google search volume of selected keywords related to financial markets increases before stock markets fall. They further show that trading strategies including information on search query changes yield higher returns compared to random trading strategies. Hamid and Heiden (2015) use daily and weekly Google search volume data to forecast the volatility of the Dow Jones Index based *inter alia* on the concept of empirical similarity. The model performs better than traditional models in in-sample and out-of-sample forecasting, particularly when using weekly data.

The literature discussed so far uses GSQ indices in the context of countries with high internet-adoption rates. Carrière-Swallow and Labbé (2013) and Seabold and Coppola (2015) are, to our knowledge, the first studies to use GSQ indices in contexts associated with significantly lower internet adoption rates. Carrière-Swallow and Labbé (2013) develop a GSQ index of online interest in automobile purchases in Chile to now-cast automobile sales. They test the now-casting capacity by comparing a benchmark model to a GSQ-augmented model. They find that models including the GSQ index can outperform benchmark models in in- and out-of-sample forecasts. Seabold and Coppola (2015) use a GSQ index to forecast aggregate consumer prices and a selection of staple food prices (beans, maize, rice, wheat, and soy) in Costa Rica, El Salvador and Honduras. Similarly to Carrière-Swallow and Labbé (2013), they use an out-of-sample estimation scheme to test the now-casting capacity of GSQ-models and non-GSQ benchmark models. They were partially successful in improving now-casts of food prices and indicate that the food price crisis of 2007/08 could be one driver, which complicates food price forecasts.

This overview shows that GSQ data has been successfully used in a variety of disciplines, while few analyses have linked Google Trends to a developing country context or food price monitoring and disaster early warning. Therefore, our contribution to the literature is threefold: (1) we are the first study to use GSQ data in an African context, (2) to analyze a larger country panel, (3) to explicitly link GSQ data to food security and to add to the knowledge on how citizen science can help to improve early prediction of food insecurity and crises.

3. Sample Considerations of Internet Data in Africa

When analyzing data derived from the Internet in a developing country context, the underlying sample characteristics are, to a large extent, unknown. This is due to a general lack of comprehensive, disaggregated end-user and infrastructure statistics across Africa. This is even more evident in the case of Google data. No information on the sample characteristics is available, as Google generally does not publish information about its end-users and their search history due to privacy concerns. Nevertheless, the following is an attempt to approximate the sample characteristics of Internet data in the underlying nine countries, by investigating the spatial spread of certain infrastructure, on which Internet access, to some extent, depends and by extrapolating from market developments on other continents.

Internet users rates across Africa are comparatively low, when compared to the rest of the world (World Bank Group 2016). Figure 1 shows the development of Internet user rates in the nine countries underlying this study. Internet user rates started to increase significantly between 2007 and 2010 and (more than) doubled in the nine countries between 2010 and 2016. As of 2016, Zambia ranks as the country with the highest Internet user rate (25%), followed by Zimbabwe and Uganda. In Rwanda, Mozambique, Kenya and Ethiopia between 15-20% of the population use the Internet, while Tanzania and Malawi rank at 10-15%.

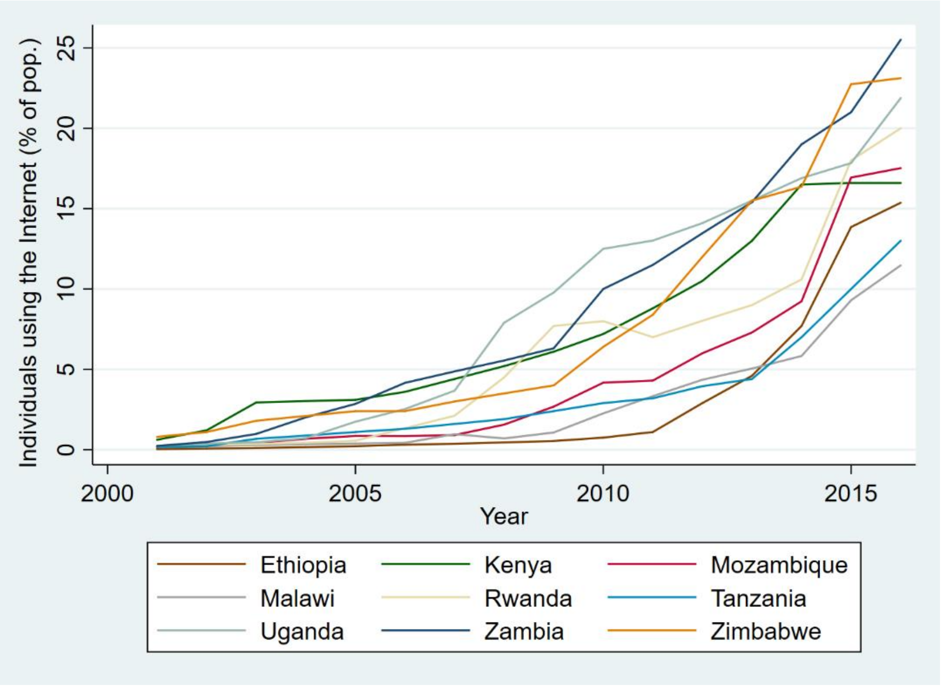


Figure 1: Internet User Rates in the 9 Study Countries.

Note: Internet users refers to individuals that have used the Internet from any location and device in the last three months. Source: Own compilation based on data from The World Bank (2018).

Given Africa's extensive landmass and limited purchasing power of consumers, the provision of cable-dependent broadband Internet is not cost effective. This is why mobile data and smart-phone adoption play a significant role in accessing the Internet and much of the increase in Internet adoption rates has been driven by mobile Internet subscribers (GSMA 2018). Due to this fact, we refrain from using the distribution of (optic fiber) cables as proxy location for Internet users. In that regard, electricity is a predominant feature necessary to access the Internet. People require a connection to an electricity grid, either to charge their mobile devices or to power their computers and modems. We hypothesize, that the availability of electricity correlates with population density. This correlation is visualized in Figure 2, where we plot the population density per km² in the nine African countries, as well as the available electricity grid. The map underlines the previous hypothesis, demonstrating that electricity grids are more prominent in urban areas and regions associated with higher population density. This indicates that data derived from the Internet is biased towards urban areas.

Apart from the basic infrastructure, that is required to go online, also socio-economic aspects drive access to Internet. A digital divide is not only prominent across countries, but also within countries, as access to Internet depends on education, literacy and income levels, as well as age (Poushter, Bishop, and Chwe 2018). For example, GSMA (2018) states that affordability of mobile services will be the major challenge in the upcoming years, with respect to increasing mobile broadband use in Africa. Across many African countries also a gender gap is still prominent, with men having more access to the Internet than women (Poushter, Bishop, and Chwe 2018). Furthermore, Weidmann et al. (2016) highlight that ethnicity plays a role in infrastructure provision, such as the expansion of the cellular network or the electricity grid, showing that different ethnic groups are discriminated with regards to Internet supply. These factors lead us to conclude that our sample is biased towards urban areas and is driven, to some extent, by younger end-users with a higher education level, who are more likely to be male. We acknowledge that the sample is non-representative of the society at large.

We further hypothesize that the sample characteristics of Internet data are not constant over time. This study covers an eleven year time span from 2006 to 2017, a period that has been marked by significant growth in Internet user rates in the nine countries. This in turn has consequences for the sample composition. Even though Google standardizes its search-query data to remove any trend stemming from increasing Internet use, we hypothesize that the sample composition did change over the study period. As Internet provision, mobile data and devices have become significantly cheaper over time (GSMA 2018), the Internet has become more accessible to a wider range of people and, consequently, more inclusive. We assume that this has been particularly the case after the year 2010, the point from which Internet user rates started to increase significantly (see Figure 1). Given that the sample is non-representative of the society and likely changes over the study period, we do not use GSQ data to predict the price level at a given time, but make use of changes in GSQ volume to detect abnormalities in food prices which may hint at upcoming crises.

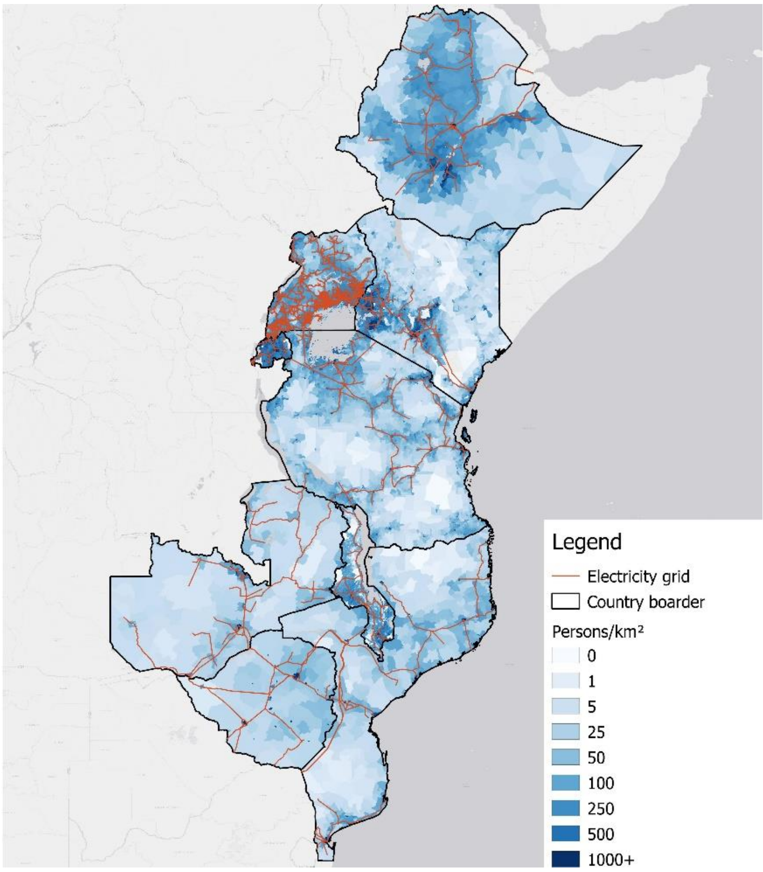


Figure 2: Population Density and Electricity Grids.

Note: Africa Electricity Grids Explorer (2017) states that the here shown electro-grids data is the most comprehensive and up-to-date public data set available for Africa. Nevertheless, the data is part of an ongoing mapping initiative and maybe, to some extent, outdated and should be used for illustration purposes only. Source: Own cartography based on population data by Center for International Earth Science Information Network Columbia University (2017), and electro-grids data by Africa Electricity Grids Explorer (2017).

After narrowing the sample characteristics further down to the average composition of the present sample, the question arises who is interested in information acquisition about food commodities. We hypothesize that these could be farmers, growing and selling their crops, traders, interested in buying and selling commodities, as well as financial institutions, insurers, governmental institutions, NGOs, international organizations, researchers and the public interested in monitoring the market. As some of these actors, *inter alia*, represent the supply and demand side of the market and as prices are a function of supply and demand, *P=f(S,D)*, we assume that GSQ data could have the potential to capture a contemporaneous price signal.

After having hypothesized about the sample characteristics of Internet data in a developing country context, we now continue to outline, why Google's search engine data can be considered a valid sample of the population with Internet access. Even though there is a lack of credible and accessible data on Google's share in the African search engine market, we assume that Google has a dominant role in Africa given the following aspects. First, its search engine market share exceeds 90% in most European countries (The Economist 2017). Second, Google's global market share ranges at 59% and its dominance is even larger in the mobile and tablet devices market, where the market share is 90.8% (Bulut 2017). Android-based smartphones and tablets, i.e. devices with an operating system developed by and based on Google, are dominant across Africa. For instance, GSMA (2018) reports that Samsung devices are still the leading player in the African mobile device market. These devices are Android-based, which means that Google Chrome is the pre-installed browser and, hence, Google is the default search engine. Due to these aspects, we assume that Google search volume is a representative sample of the population that uses the Internet. Hence, we do not assume that the presence of other search engines introduces further bias in our sample.

4. Data

As data availability and quality are major limitations across African countries, we use a data driven approach to select the countries for our analysis. We include all countries, in which maize plays an important role as staple crop in the country's food basket and a sufficient amount of data is available. This refers to monthly agricultural price data and GSQ data with a sufficient search volume. We include Ethiopia, Kenya, Malawi, Mozambique, Rwanda, Tanzania, Uganda, Zambia and Zimbabwe in our analysis. Due to data constraints regarding food prices, the time period for the analysis ranges from 01.2006 to 07.2017. GSQ data are generally available since 2004.

We use monthly staple food price data as provided by FAO GIEWS. The data availability of food prices varies across countries. We retrieve maize prices in nominal US Dollar/ton at the respective capital markets (in the case of Tanzania, we download maize prices for Dara salaam). Nominal prices in USD, different to real prices in USD and nominal prices in local currency units, do not contain noise from general food price inflation. Due to many missing variables in the maize price series of Malawi and Zimbabwe, we retrieve maize prices for the two countries from the ZEF price database. We use simple linear interpolation in case of missing observations.

We download monthly GSQ data from Google Trends, https://trends.google.com, as this matches the frequency of the maize price data. Google Trends provides an index of search activity for a specified search word at a given location and point in time. The index is a measure of the relative popularity of one search term as a fraction over the total body of search volume, since Google does not publish its absolute search volume. GSQ data is further being transformed by two steps prior to publication: the index is normalized, meaning that it is divided by total search queries in a given location at a specific point in time. This normalization removes any trend from the data that could stem from growth in Internet users or changes in Google's popularity as a search engine (Carrière-Swallow and Labbé 2013). It is also standardized, as it is scaled from 1 to 100 and averaged to the nearest integer.

There are a variety of challenges and particularities associated with GSQ data, which have strong implications for the analysis and data sampling: Firstly, GSQ data is a relative index. When comparing two series to each other, one particularly popular series might push the more unpopular series towards zero. To overcome this issue, we downloaded each series separately for each country by restricting the geographical unit. When downloading the series separately, we lose the ability to compare the normalized series to each other, which leaves us with an analysis of growth rates across series.

Secondly, Google changes its data provision. At two points within the sampling period, Google implemented changes to the data, noting that on 01/01/2011 “an improvement to our geographical assignment was applied” and on 01/01/2016 “an improvement to our data collection system was applied" (Google 2018b). Google does not give any further information on the adjustment procedure. Hence, these changes in the data cannot be explicitly taken into account in the analysis.

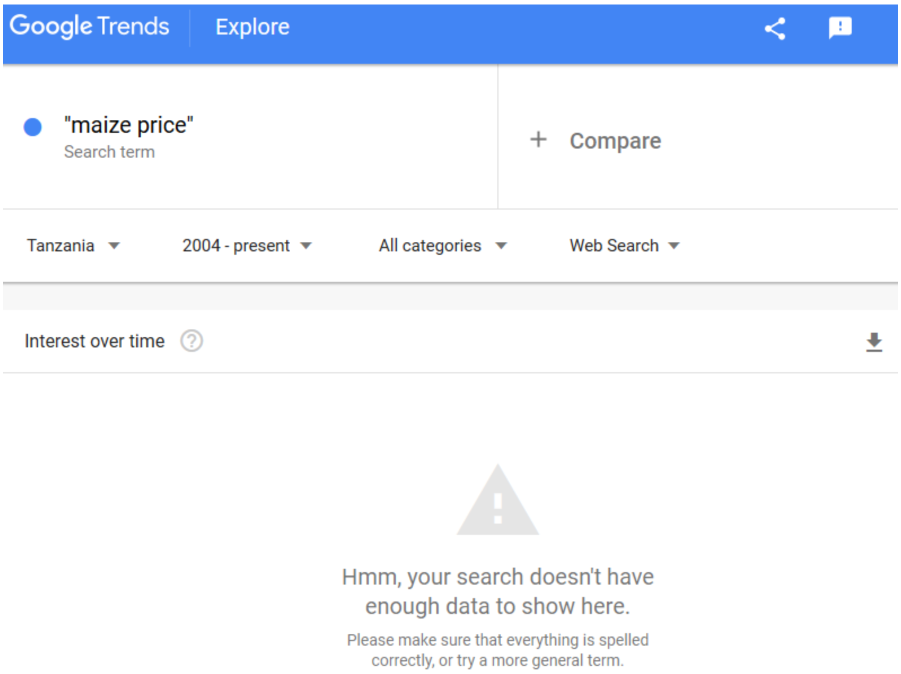
Thirdly, Google Trends has an unreported privacy threshold. This means that the search index is only reported in case the search volume exceeds a specific threshold, which is based on absolute search volume and unknown to the public. If the threshold is not passed, the search volume is automatically reported as zero (Stephens-Davidowitz and Varian 2014). The observance of zero values is problematic, as we do not observe a signal, where, theoretically, there should be one. The fact that search volume is only reported after passing an unknown threshold is particularly problematic in developing countries, where the search volume is generally lower, given that there are lower internet-adoption rates and, hence, less signal producing users. When downloading the data for African countries, we observe a large occurrence of zero values. It is unknown, whether Google has different privacy thresholds for different countries. This threshold is further the reason, why we choose country-level data for the analysis. Currently, Google Trends provides data at the sub-regional level for all analyzed countries, with Kenya being the only exemption. The sub-regional search volume, however, is still very low. Hence, we observe a very large amount of zero search volume or no search volume is reported at all for location specific data. Thus, we follow Stephens-Davidowitz and Varian (2014) and downloaded the data for a coarser geographic unit, i.e. at the country level.

Fourthly, GSQ data is unstable over time. This means that downloading the same sample on different days yields a different time series of search volume. The data, however, remains stable within the same 24 hour period. This is due to Google drawing the single, requested sample from its absolute body of search volume, while Google seems to cache its data daily. This is why the same sample request remains the same over 24 hours (see Stephens-Davidowitz, 2013; Seabold and Coppola, 2015; Carrière-Swallow and Labbé, 2013). To deal with this instability of data over time, previous studies chose to draw samples of the same search query over a longer time period as an attempt to approximate the "true" Google search volume. Carrière-Swallow and Labbé (2013), for example, downloaded GSQ data on 50 occasions, while Seabold and Coppola (2015) drew samples on 10 days within one month.[1]

To choose potential search terms or predictors, Stephens-Davidowitz and Varian (2014), Scott and Varian (2013) and Lazer et al. (2014) highlight the importance of using variable selection techniques instead of simple judgment. This is to achieve a better model fit and to avoid so called "fat-regression" problems, i.e. models where the number of possible predictors exceeds the number of observations. These studies, however, use Google Trends in a non-developing country context and rely to a large extent on Google Correlate. Google Correlate is an online tool, where one can upload a given time series and it will provide a ranking of search-term series depending on the degree of correlation between the two series (Google 2018a). At the

[1] See below for a detailed description of our methodology.

time of this study, Google Correlate is unavailable for the study countries. Hence, we proceed with simple judgment regarding the selection of Google search terms and choose the most parsimonious keyword, i.e. *maize*. This keyword was chosen *ex-ante* (1) due to the belief that it contains relevant information that will allow us to use it as a proxy for price developments and (2) due to Google's privacy threshold, which has not only consequences for the choice of geographical unit, but also influences the choice of search terms. Any potential and more precise combination of words, like "*maize price*", frequently pushes the search volume below its reporting threshold and, hence, defaults to not being reported. This scenario can be seen in Figure 3, which shows that the search volume for the term "*maize price*" in Tanzania does not exceed the privacy threshold and is, consequently, not reported.



"maize price"
Search term

Tanzania ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Hmm, your search doesn't have enough data to show here.

Please make sure that everything is spelled correctly, or try a more general term.

Figure 3: Privacy Threshold and Search Term Choice in Tanzania.

Source: Screenshot taken from https://trends.google.com on Nov, 16th 2018.

Furthermore, this study deals with nine different countries, where different languages are being spoken. To choose the language of search terms for each country, we compare the search volume of the English keyword, to the search volume in the respective national language, for example Kiswahili in Kenya and Luganda in Uganda, with the aim to understand how Internet users interact with Google. The direct comparison of search terms needs to be performed within the Google Trends tool, as this is the only way to ensure the comparability of search volume across keywords at a given point in time and spatial unit. An exemplary illustration of this comparison of search terms in English *maize* and the official language *kasooli* for the case of Uganda can be found in the appendix. We find that for all countries the volume of English keywords exceeds the volume of keywords in other official languages and, hence, proceed by using the English search term. By doing so, we follow other studies like

Almanzar and Torero (2017) who compare the Google News Feed in English to the local language and also opt for English search terms.

After delineating the search term and language choice, we now address the above discussed instability of GSQ data. To approximate the “true” GSQ value, we follow Seabold and Coppola (2015) and Carrière-Swallow and Labbé (2013) and draw samples of each data series of each country on 30 different days. We calculate the “true” GSQ value as the mean of 30 samples, which we will continue to use in the analysis.[2] Figure 4 illustrates the maximum and minimum GSQ value observed for the search word *maize* within the 30 samples, as well as the calculated mean GSQ value. For illustration purposes we show the data for Ethiopia, Kenya, Uganda and Tanzania (Figures for the remaining five countries can be found in the appendix). We can see that the variation in the sample reduces significantly post 2011, which coincides with Google's "improvement of geographical assignment". We further see that we draw many samples with zero search volume. The incidents of zero search volume, however, could be reduced significantly by averaging over the samples and we observe few observations where the search volume is zero at mean, which is still the case particularly in earlier periods of the series. This reduction in zero observations leads us to assume that we are able to approximate the “true” signal by the repeated sampling of GSQ data.

In Figure 5, we plot the development of the mean GSQ value of the keyword *maize* as well as maize prices in the same countries. We observe that the GSQ data is generally more volatile than the maize price series. In all countries, an increase in maize price around the food price crises of 2007/08 and 2011/12 is visible. We further note that spikes in GSQ data coincide with spikes in maize price data, which is particularly visible in the case of Kenya around 2010, 2012 and 2018.[3]

[2] The underlying data is available upon request.

[3] Summary statistics of each series can be found in Table A of the Appendix.

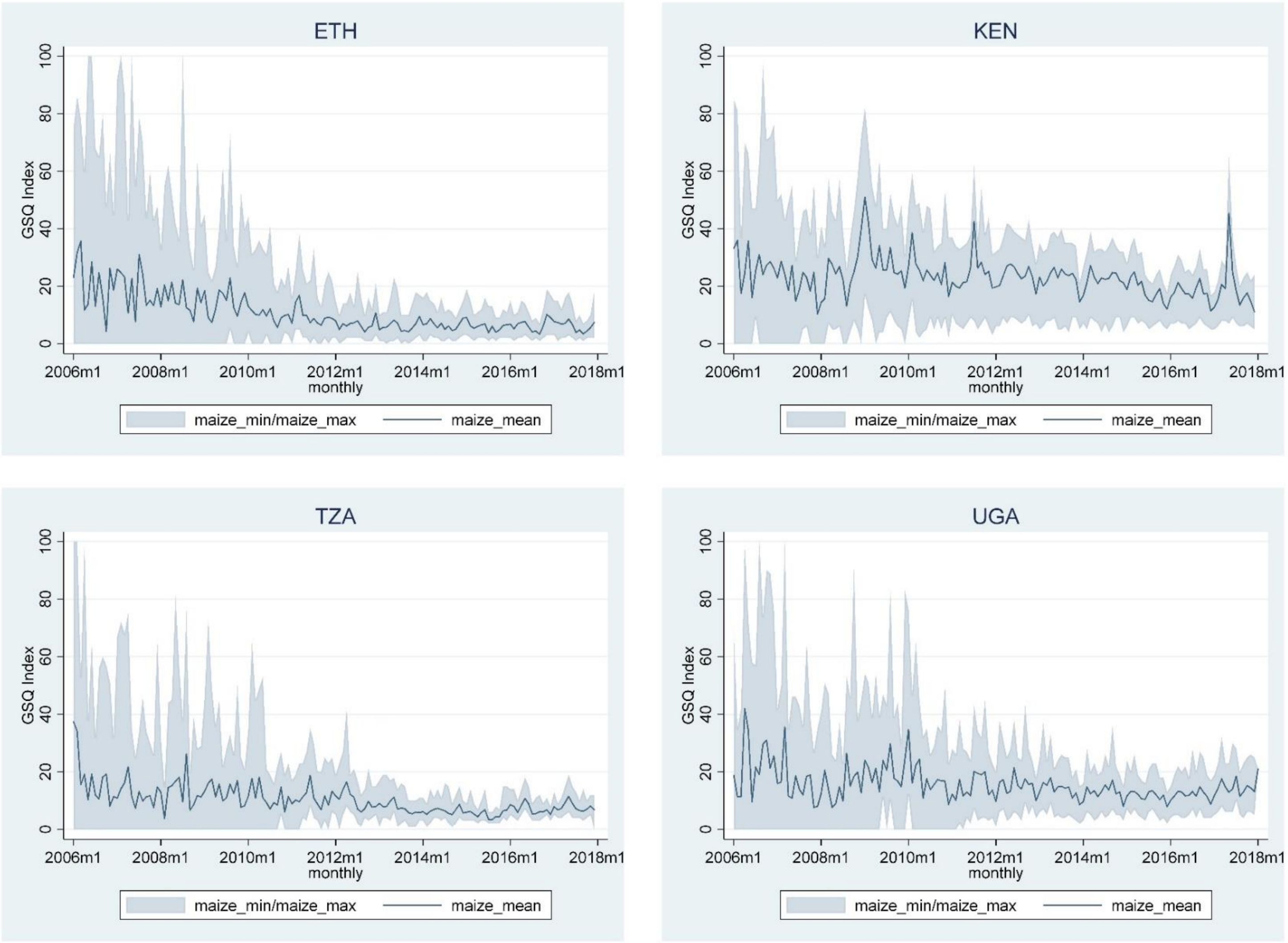


Figure 4: Sampling Noise of GSQ Data for the Search Term *maize* in Ethiopia, Kenya, Tanzania and Uganda.

Source: Own compilation based on data extracted from https://trends.google.com, sampled over a period of 30 days in December 2017.

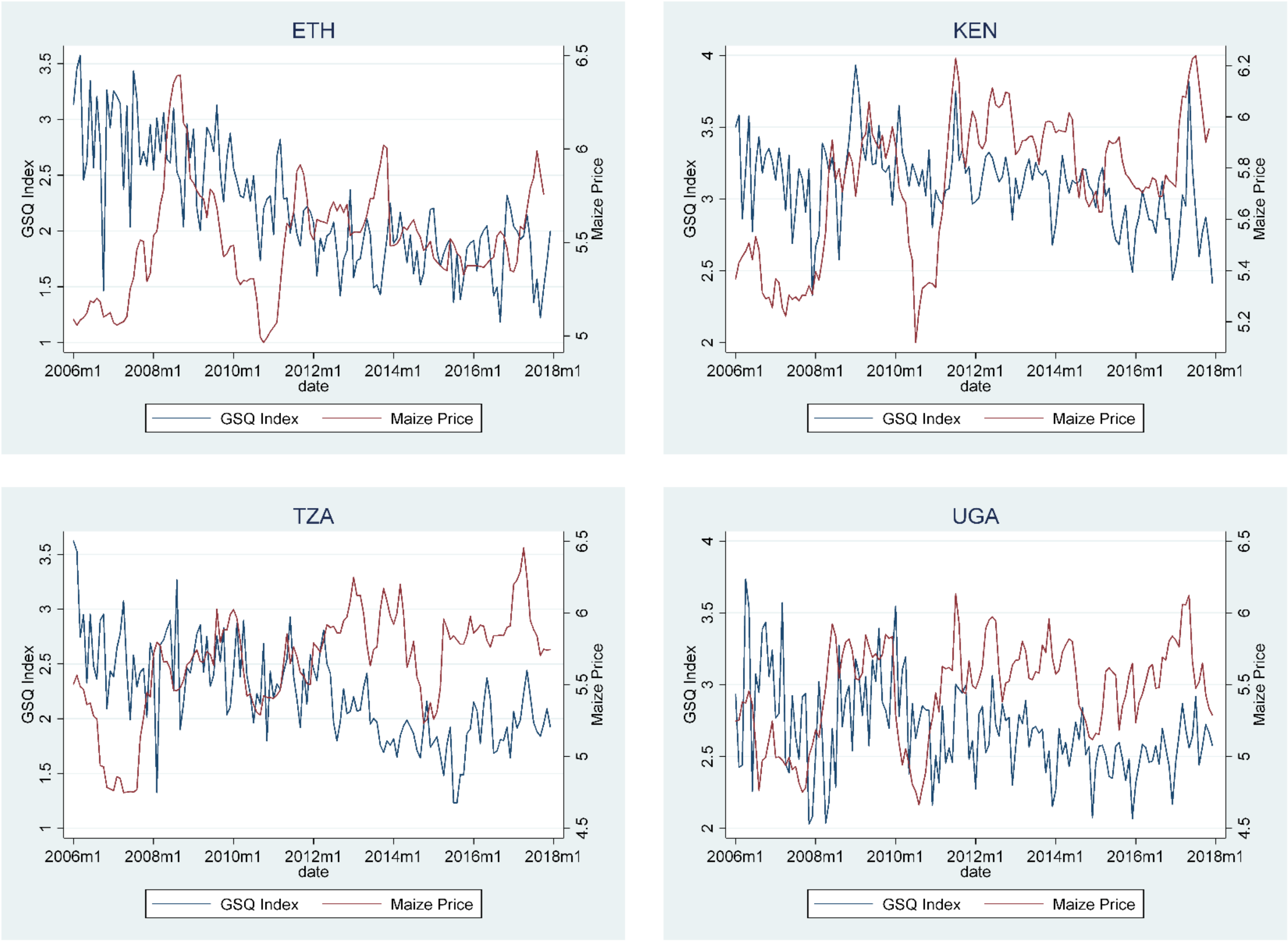


Figure 5: GSQ Data for the Search Term *maize* and Maize Prices in Ethiopia, Kenya, Tanzania and Uganda.

Source: Own compilation based on data extracted from https://trends.google.com and FAO GIEWS.

5. Methodology

To test whether GSQ data can contribute to the now-casting of maize prices in selected African countries, we pursue a two-tiered estimation strategy. We start with the in-sample estimation of a benchmark and a competing, GSQ-augmented, model for each country. We subsequently continue with the evaluation of the two competing models in a pseudo out-of-sample forecasting environment to test the now-casting performance of the two specifications based on their out-of-sample forecasting errors.

Before the estimation, we inspect the respective data series with regards to its time series properties. After replacing missing values in the price series by simple linear interpolation and logarithmizing both price and GSQ data, we assess the order of integration of each series using Philips-Perron unit root tests (Philips-Perron test statistics are reported in the appendix). In the case of Kenya and Zimbabwe, the respective series are non-stationary and we proceed with first differences, given as $\Delta Y_t = y_t - y_{t-1}$, where ΔY_t is the change of Y between periods t and $t-1$.

5.1 In-Sample Estimation

To analyze whether GSQ data improves the now-casting accuracy of maize prices, we follow Choi and Varian (2012) as well as Carrière-Swallow and Labbé (2013) and formulate a benchmark model and two competing, GSQ-augmented models. The objective of this study is to forecast the present. Assessing the in-sample fit of models is not sufficient to draw conclusions about a model's forecasting ability, due to issues of over-fitting and data mining, as well as the potentially large differences in model fit of in-sample prediction and out-of-sample forecast (Stock and Watson 2015). This is why we use the in-sample estimation solely (1) to understand the relationship between maize prices and GSQ data and (2) to show that the benchmark model is an appropriate specification, given that a causal interpretation is irrelevant for forecasting, as the focus is on a predictor's ability to improve a model's forecasting capacity and not on causality (Stock and Watson 2015).

As benchmark, we fit simple, linear auto-regressive (AR) models to the maize price series y in each country i. We determine the optimal number of lags of the dependent variable y_i based on the Schwarz's Bayesian information criterion (SBIC) (test statistics are reported in the appendix). As both series exhibit a degree of seasonality, we control for the presence of deterministic seasonality by including monthly dummy variables. As benchmark, we estimate

$$y_{i,t} = \alpha_{a,i} + \sum_{k=1}^{p} \beta_{a_{i,k}} y_{i,t-k} + \sum_{j=1}^{s-1} \gamma_{a_{i,j}} D_{j,t} + \varepsilon_{a_{i,t}} \qquad \text{Eq. (1)}$$

where $y_{i,t}$ is the maize price in country i and time t, $t-k$ is the optimal number of lags of the dependent variable of country i based on the SBIC, $D_{i,t}$ are the seasonal dummy variables with $s = 12$ and $\varepsilon_{a_{i,t}}$ the white noise error term.

We estimate the following GSQ-augmented model, which we augment by adding the contemporaneous GSQ value, GSQ(1), for each country i

$$y_{i,t} = \alpha_{b,i} + \sum_{k=1}^{p} \beta_{b_{i,k}} y_{i,t-k} + \sum_{j=1}^{s-1} \gamma_{b_{i,j}} D_{j,t} + \delta_{b,i} GSQ_{i,t} + \varepsilon_{b_{i,t}} \qquad \text{Eq. (2)}$$

where $y_{i,t}$ is the maize price in country i and time t, $GSQ_{i,t}$ is the Google keyword *maize* in country i and time t; $t-k$ is the optimal number of lags of the dependent variable of country i based on the SBIC and $\varepsilon_{b_{i,t}}$ the white noise error term.

Moreover, we hypothesize that the value of GSQ_t depends on the direction of the change in maize prices. Thus, to further dis-entangle the relationship between maize price developments and GSQ data, we estimate a second GSQ-augmented model, GSQ(2), in which we interact GSQ_t with a dummy variable that equals 1 for a positive maize price change:

$$y_{i,t} = \alpha_{c,i} + \sum_{k=1}^{p} \beta_{c_{i,k}} y_{i,t-k} + \sum_{j=1}^{s-1} \gamma_{c_{i,j}} D_{j,t} + \delta_{c,i} GSQ_{i,t} + \zeta_i GSQ_{i,t} x\, Z_{\Delta Y_i} + \varepsilon_{c_{i,t}} \qquad \text{Eq. (3)}$$

where $y_{i,t}$ is the maize prices in country i and time t, $GSQ_{i,t}$ is the Google keyword *maize* in country i and time t; $t-k$ is the optimal number of lags of the dependent variable of country i based on the SBIC, $GSQ_{i,t} x\, Z_{\Delta Y_i}$, is the interaction term based on the current GSQ value and the dummy variable Z, with $Z = 1$ if $\Delta Y_i > 0$ and $Z = 0$ if $\Delta Y_i < 0$ and $\varepsilon_{c_{i,t}}$ the white noise error term.

5.2 Out-Of-Sample Estimation

After assessing the in-sample properties, we continue with the evaluation of the out-of-sample forecasting performance of the competing models. The objective is to understand whether contemporaneous GSQ data contains information that improves the now-casting accuracy of regular, auto-regressive maize price models. The forecasting accuracy of different models is tested in a pseudo out-of-sample context by restricting the number of observations and re-estimating the model for the remaining time periods.

Again we follow Carrière-Swallow and Labbé (2013) and Seabold and Coppola (2015) to estimate a linear, static, one-step-ahead model that is based on a recursive window scheme. We choose a static model and a recursive estimation scheme, as we anticipate this to be the scenario decision makers would engage in. The recursive window implies that the actual observation is added for each estimation period. It hence is similar to a scenario in which decision makers would add variables to their model once they become available.

We restrict the full sample to a training section and a pseudo-out-of-sample section, to test the forecasting accuracy against observed values. Under the recursive scheme, we begin by estimating the models over the first S periods of time. These estimates are then used to formulate the first out-of-sample now-cast for period $S + 1$. We then re-estimate the model for each time period by extending the estimation window forward until the end date $t \in (S + 1, \ldots, T + 1)$, where $T + 1$ is the last period in the full sample. We chose the window size S to be 36, corresponding to a time frame from January 2006 to December 2008. Hence, the forecast starts at 12.2008, which leaves us with 108 forecasted values to assess the forecasting accuracy of the competing models.

We subsequently evaluate the out-of-sample forecasting performance of the two specifications by calculating the Mean Squared Forecast Error (MSE). Following Stock and Watson (2015), the one-step-ahead forecast error of each model i is given by

$$\hat{e}_{i,t+1} = y_{i_{t+1}} - \hat{y}_{i,t+1\,|T} \qquad \text{Eq. (4)}$$

where $y_{i_{t+1}}$the observed value in country i and $\hat{y}_{i,t+1\,|T}$ the forecast of model i, estimated using observed data through time T. The MSE of country i and follows as

$$MSE_i = \frac{1}{N_T} \sum_{t=1}^{N_T} (\hat{y}_{i,t} - y_{i,t})^2 \qquad \text{Eq. (5)}$$

where N_T is the total number of observed time periods in the out-of-sample window. The model associated with the smaller MSE beats the competing model.

6. Results

6.1 In-Sample Estimation

The results of the in-sample estimation are reported in Table 1, Table 2 and Table 3, where we show the results of the benchmark estimation (Eq. 1) and the two GSQ specifications (Eq. 2 and Eq. 3). Based on the SBIC, we estimate a parsimonious AR(1) for seven out of nine countries, while we estimate a AR(2) specifications in the case of Ethiopia and Zimbabwe. When considering the R^2, the parsimonious AR specifications prove to be a good fit in the majority of countries (0.99). This is, as expected, due to the highly auto-regressive nature of price series. We achieve the lowest fit for Kenya and Zimbabwe.

When considering the first GSQ-augmented model, GSQ(1), in Table 2, we find the contemporaneous GSQ-value, GSQ_t, to be significant in four of the analyzed countries. These are Rwanda, Uganda, Zambia and Zimbabwe. We can reject the null hypothesis that the coefficient of GSQ_t is equal to zero at the 5% significance level in the case of Zambia and at the 10% level for Rwanda, Uganda and Zimbabwe. The estimated coefficients are negative for Rwanda, Uganda and Zambia, indicating that an increase in maize prices is associated with a decrease in search volume of the term maize. We observe a positive coefficient in the case of Zimbabwe. When further disaggregating the effect of GSQ_t in a second GSQ-augmented model, GSQ(2), Table 3, we find the interaction term, interacting GSQ_t with a dummy for positive price change, to be positive and significant at the 1% level in the 9 countries. These results indicate a positive relationship between maize prices and search volume, when allowing for a different slope in case of a positive and negative price change from period $t-1$ to t. However, in the case of Rwanda, Uganda and Zambia, for which we found a negative effect of GSQ_t in GSQ(1), the positive and negative coefficients are close to offsetting each other. This implies that a reduction in GSQ_t is associated with decreasing prices, but an increase in GSQ_t is not associated with higher maize prices.

Variables	ETH	KEN	MOZ	MWI	RWA	TZA	UGA	ZMB	ZWE
Maize Price (y_{t-1})	1.176***	0.155*	0.896***	0.910***	0.911***	0.937***	0.900***	0.940***	-0.350***
	(0.0913)	(0.0919)	(0.0490)	(0.0299)	(0.0368)	(0.0313)	(0.0372)	(0.0341)	(0.121)
Maize Price (y_{t-2})	-0.241***								-0.304**
	(0.0885)								(0.144)
Seasonal Dummy	yes	yes	yes	yes	yes	yes	yes	yes	yes
N	140	141	143	143	143	143	143	143	134
R^2	0.9998	0.1574	0.9997	0.9905	0.9997	0.9995	0.9993	0.9999	0.2122

Table 1: In-Sample Estimation, Benchmark Model.

Note: Robust standard errors in parentheses. ***p < 0.01, ** p <0.05, *p < 0.1. Seasonal dummy variables omitted for brevity. Source: Own estimation.

Variables	ETH	KEN	MOZ	MWI	RWA	TZA	UGA	ZMB	ZWE
Maize Price (y_{t-1})	1.176***	0.153*	0.904***	0.912***	0.901***	0.919***	0.900***	0.948***	-0.376***
	(0.0915)	(0.0913)	(0.0559)	(0.0292)	(0.0405)	(0.0348)	(0.0371)	(0.0333)	(0.115)
Maize Price (y_{t-2})	-0.240***								-0.289**
	(0.0892)								(0.133)
GSQ_t	0.00168	0.0241	-0.00501	0.0111	-0.0366*	-0.0447	-0.0736*	-0.0280**	0.161*
	(0.0126)	(0.0276)	(0.0100)	(0.0160)	(0.0214)	(0.0363)	(0.0406)	(0.0113)	(0.0891)
Seasonal Dummy	yes	yes	yes	yes	yes	yes	yes	yes	yes
N	140	141	134	142	134	143	143	143	134
R^2	0.9998	0.1615	0.9997	0.9909	0.9998	0.9995	0.9994	0.9999	0.2649

Table 2: In-Sample Estimation, GSQ-Augmented Model (1).

Note: Robust standard errors in parentheses. ***p < 0.01, ** p <0.05, *p < 0.1. Seasonal dummy variables omitted for brevity. Source: Own estimation.

Variables	ETH	KEN	MOZ	MWI	RWA	TZA	UGA	ZMB	ZWE
Maize Price (y_{t-1})	1.086***	0.0654	0.931***	0.937***	0.959***	0.946***	0.941***	0.938***	-0.359***
	(0.0628)	(0.0509)	(0.0526)	(0.0232)	(0.0283)	(0.0255)	(0.0267)	(0.0247)	(0.0766)
Maize Price (y_{t-2})	-0.108*								-0.259***
	(0.0602)								(0.0976)
GSQ_t	-0.0375***	-0.00543	-0.0329***	-0.0110	-0.0691***	-0.0468	-0.0850***	-0.0434***	0.0640
	(0.0112)	(0.0166)	(0.0106)	(0.0135)	(0.0179)	(0.0289)	(0.0317)	(0.00986)	(0.0722)
$GSQ_t\ x\ D\Delta y$ $(1\ if\ \Delta y > 0)$	0.0527***	0.0418***	0.0805***	0.0729***	0.0691***	0.0807***	0.0798***	0.0417***	0.138***
	(0.00553)	(0.00341)	(0.0125)	(0.00889)	(0.00713)	(0.00773)	(0.00744)	(0.00389)	(0.0233)
Seasonal Dummy	yes	yes	yes	yes	yes	yes	yes	yes	yes
N	140	141	134	142	134	143	143	143	134
R^2	0.9499	0.6003	0.8352	0.8842	0.9211	0.9424	0.9105	0.9482	0.5348

Table 3: In-Sample Estimation, GSQ-Augmented Model (2).

Note: Robust standard errors in parentheses. ***p < 0.01, ** p <0.05, *p < 0.1. Seasonal dummy variables omitted for brevity. Source: Own estimation.

6.2 Out-Of-Sample Estimation

We continue with the evaluation of the now-casting performance of the competing models. In Table 4 we report the MSE of the one-step-ahead out-of-sample now-cast of the benchmark and the two GSQ-augmented models. We observe that the MSE of the benchmark specification is relatively low, indicating that past price observations are a good basis to forecast maize prices and that the estimated AR models perform well also in out-of-sample forecasts.

When comparing the MSE of the benchmark with the first GSQ-augmented model, we achieve a reduction in MSE in 7 out of 9 countries. We obtain the largest improvement of MSE in the case of Zambia and Tanzania, with an improvement in forecasting fit by 14.95% and 14.23% respectively. This is followed by Uganda, Rwanda, Kenya and Mozambique, where improvements range between 3% and 8%. Also the forecast for Malawi could be improved, if marginally, by 0.82%. In the case of Ethiopia and Zimbabwe, the GSQ-specification yields larger errors and, hence, a reduction in forecasting fit. Particularly in the case of Zimbabwe, we observe a large increase in MSE. In summary, the first GSQ specification, GSQ(1) beats the benchmark model at mean in 7 out of 9 countries and including contemporary GSQ data improves the fit of maize price now-casts.

When comparing the forecasting errors of the second GSQ-augmented model to the benchmark model, we find an improved forecast fit in 4 out of 9 countries. These are Malawi, Kenya, Tanzania and Ethiopia, with a reduction in MSE by 23.41%, 17%, 5.29% and 3.62% respectively. For Malawi, Kenya and Ethiopia, the second GSQ specification also provides the smaller MSE when compared to the first GSQ specification, which is not the case for Tanzania, where the first GSQ specification yields better now-casts. Overall, the first GSQ-augmented model achieves an improved now-cast in more countries, when compared to the second GSQ specification. This might be due to the fact that interaction terms tend to be variations of already included information in the forecasting model and hence lead to imprecision in an out-of-sample, forecasting setting (Lindh 2011). Still, in the case of Ethiopia, Kenya and Malawi, GSQ(2) provides the better forecast fit. When considering both GSQ specifications, we achieve an improvement in forecasting fit in 8 of 9 countries. Hence, by including contemporaneous search engine metadata, we improve the now-casting capacity of simple AR models that are based on past price realizations. The only exemption is Zimbabwe, where the benchmark model beats both GSQ-augmented models.

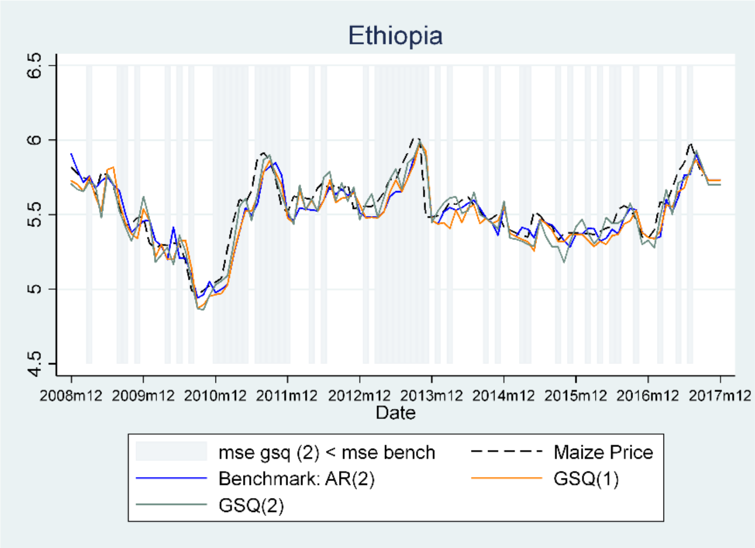
Figure 6 shows the results of the one-step-ahead out-of-sample forecast for Ethiopia, Tanzania and Uganda, for which we find the GSQ-augmented models to beat the benchmark model at mean (see appendix for remaining figures). We can see that forecasted values of the benchmark and GSQ-augmented model follow the actual maize price movements, illustrating the low MSE of the competing models. Following Choi and Varian (2012), we display the forecasting error over time and show in which instances the GSQ-augmented model beats the benchmark model (grey shading).

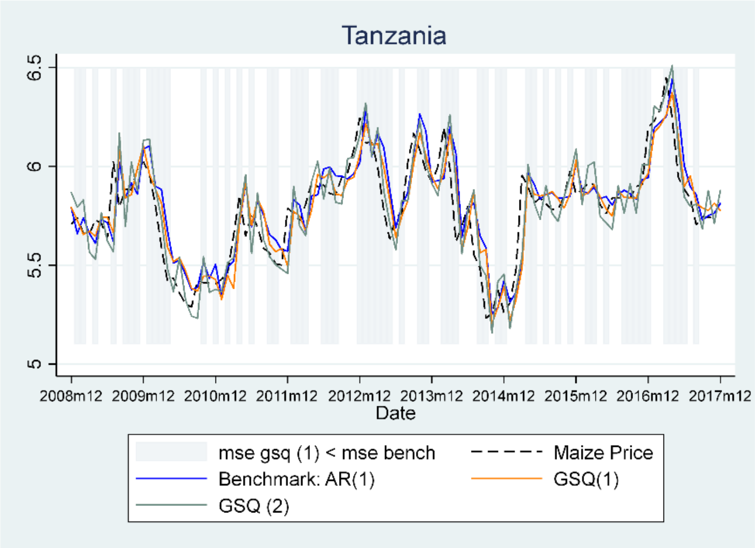
Country	Benchmark	GSQ (1)	GSQ (2)	BM vs GSQ(1)	BM vs GSQ(2)
ETH	0.011801	0.013310	0.011373	12.79	-3.62
KEN	0.011808	0.011241	0.009801	-4.80	-17.00
MOZ	0.015921	0.015435	0.016086	-3.05	1.04
MWI	0.034375	0.034094	0.026328	-0.82	-23.41
RWA	0.012270	0.011441	0.014866	-6.76	21.16
TZA	0.022633	0.019413	0.021436	-14.23	-5.29
UGA	0.032399	0.029940	0.041347	-7.59	27.62
ZMB	0.009790	0.008327	0.010371	-14.95	5.93
ZWE	0.066244	0.119440	0.136219	80.30	105.63

Table 4: MSE of One-Step-Ahead Forecast, Out-of-Sample Estimation.

Note: BM=Benchmark. Source: Own estimation.

The aim is to understand whether there are certain time periods in which the GSQ-specification provides the better forecast fit. In Zambia the GSQ(1) provides the better now-cast of the increase and decline of maize prices in late 2013 and it provides a better forecast fit for the year 2017. In the case of Tanzania, we observe a cluster of smaller now-casting errors during the maize price increase and decrease around 2009, 2012 and 2016/2017. In Ethiopia, GSQ(2) seems to identify price increases and peaks better in the period from 2010 to 2012 as well as in 2013. While the second GSQ model seems to be a good specification in the case of Ethiopia, it overshoots price spikes and fails to identify price developments in the forecasts for Tanzania and Zambia, which can also observed in the remaining countries, where the first GSQ model provides better forecasts. This corroborates the point that the interaction term introduces inaccuracies in the majority of now-casts. From visual inspection, the GSQ-augmented models seem to, *inter alia*, outperform the benchmark models around peaks and turning points. It would be of interest to explore this relationship further, in particular in light of the special interest of the development community to predict those peaks and turning points.





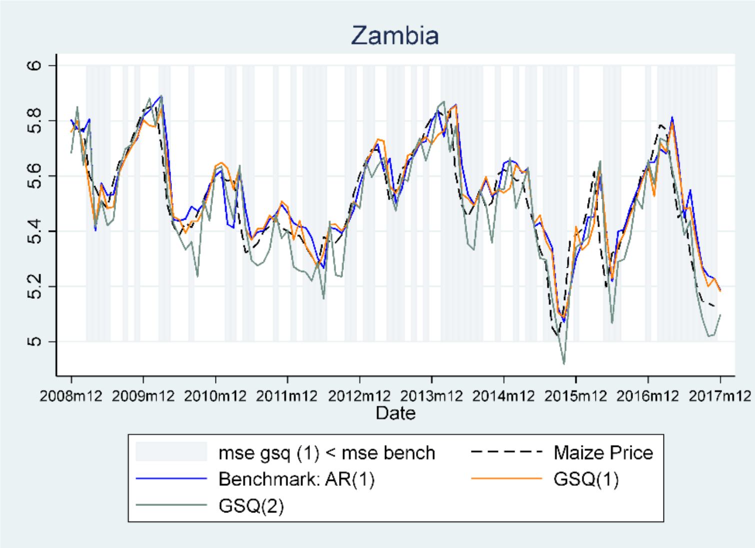


Figure 6: Benchmark vs. GSQ-Augmented Out-Of-Sample Forecasts for Ethiopia, Tanzania and Zambia.

Note: In-sample training period (01.2006 - 12.2018) not displayed. Source: Own estimation.

6.3 Discussion

In the in-sample scenario, we find the inclusion of the GSQ keyword *maize* into simple AR models for maize prices to be significant in four of the analyzed panel of nine countries. These countries are Rwanda, Uganda, Zambia and Zimbabwe. Unexpectedly, we find this relationship to be negative, i.e. an increase in maize prices is associated with a decrease in search volume. When we further dis-entangle the relationship between maize price developments and GSQ values by interacting GSQ_t with a dummy indicating a positive change in maize prices, hence, allowing for a different slope in the event of a positive price change, we find a significant and positive relationship between maize prices and GSQ values in all countries. Hence, in the majority of countries, an increase in maize prices is associated with an increase in search volume of the term *maize*.

When tested in a pseudo-out-of-sample, one-step-ahead forecasting environment, our results indicate that the GSQ-augmented models beat the benchmark model in 8 out of 9 analyzed countries. By including contemporaneous search engine data into now-casting models, we achieve a substantial improvement in forecasting fit that ranges between 3% and 23%. We achieve the largest reduction of now-casting error for Malawi, Ethiopia, Kenya, Zambia and Tanzania. Our results indicate that online signals in form of search engine metadata contain information that helps to identify maize price developments, also in environments with low Internet-adoption rates. Hence, it would be of interest to further analyze this relationship and how online signals could be systematically harnessed and integrated in forecasting and early warning models.

Zimbabwe is the only country for which the benchmark model beats both GSQ specifications. This is an unexpected finding, given that Zimbabwe has one of the highest Internet-user rates in the country panel, and, presumably, a relatively strong online signal. Also Seabold and Coppola (2015), who were partially able to improve now-casts of food prices in Costa Rica, El Salvador and Honduras by including GSQ data, contemplate potential reasons for difficulties in forecasting. They hypothesize that the complication in now-casting of food prices is likely due to the occurrence of the global food price crisis in the years 2007/08. Also in our case, the food price crisis coincides with our in-sample training period, which runs from 2006 to 2008. In the case of Zimbabwe, however, the nature of the underlying maize price series could as well drive this difficulty in forecasting, since it exhibits a strong degree of price volatility prior to 2010 and hence during our in-sample training period, followed by little to no variation in the years 2010-14 (see Figure C). During the sample period, Zimbabwe experienced multiple periods of hyperinflation, which might contribute to difficulties in taking up price signals with the GSQ series. Lastly, and not limited to the case of Zimbabwe, doubt about the quality of the maize price data as food security indicator may be well justified. In this case, GSQ data might reflect the current food security situation, but the price data does not (e.g. due to political influence on price data).

As part of this empirical exercise, our study identifies various challenges that arise when working with GSQ data in an environment with low Internet-adoption rates: Google's opaque data sampling characteristics, data instability, the relative nature of the index, and Google's manipulations of data sampling techniques without providing details, is particularly problematic. Furthermore, the unknown privacy threshold complicates analyses in environments with low Internet-adoption rates, as the signal is frequently too low to pass the reporting threshold and consequently pushes researchers to adopt coarser geographical units, data frequencies and broader search terms. Hence, valuable signal is lost. This experience might help to inform other researchers and practitioners, interested in similar research questions and contexts.

Nevertheless, the exploratory nature of our study and our study being, to our knowledge, the first attempt at using GSQ data in an African context, gives reason to further investigate the potential of GSQ data as signal for (food) price developments across Africa and other environments with low Internet-user rates. With its search engine data, Google provides a stable and cost-effective source of online signal that proves itself to be of interest for future research. Furthermore, the continuous increase in Internet-user rates across Africa will contribute to a more robust online signal in the upcoming years, which would mitigate some of the challenges that currently arise when working with GSQ data.

7. Conclusion

This study focuses on exploring the link between search engine data and food prices and analyzes the potential of search engine metadata for food price monitoring in an African context. More precisely, this analysis evaluates whether GSQ data can improve now-casting models of maize prices in nine African countries, namely Ethiopia, Kenya, Mozambique, Malawi, Rwanda, Tanzania, Uganda, Zambia and Zimbabwe.

Our study finds the inclusion of the GSQ keyword *maize* into simple AR models for maize prices in an in-sample scenario to be significant in four of the analyzed panel of countries. These are Rwanda, Uganda, Zambia and Zimbabwe. Furthermore, a specification, in which we include GSQ data as interaction term with a price change dummy, shows a significant and positive relationship, i.e. an increase in maize prices is associated with an increase of the search term *maize*, in the majority countries.

In a pseudo-out-of-sample, one-step-ahead forecasting environment, we find the GSQ-augmented models to beat the benchmark AR model in 8 of the 9 countries included in this study. Zimbabwe is the only country, for which forecasts could not be improved. By including the GSQ data, we reduce the now-casting error of maize prices between 3% and 23% and achieve the largest improvement of maize price now-casts for Malawi, Kenya, Zambia and Tanzania, with improvements larger than 14%. Our results indicate that including contemporaneous search engine data can improve the now-casting capacity of maize price models, which are solely based on past price observations.

The exploratory nature of our study gives reason to further investigate the potential of GSQ data as signal for prices developments. Future research should explore ways for the systematic harnessing and integration of online signals for forecasting and early warning models, the options of using higher frequency data, as GSQ data is potentially available at weekly frequency, more sophisticated variable selection techniques and models, like mixed frequency times series models, as well as to construct a search-query index that includes English and non-English search words and multiple crops.

References

Africa Electricity Grids Explorer. 2017. *Africa - Electricity Transmission and Distribution Grid Map*. Washington D.C.: World Bank. http://africagrid.energydata.info/.

Almanzar, Miguel, and Maximo Torero. 2017. “Media Coverage and Food Commodities: Agricultural Futures Prices and Volatility Effects.” *ZEF Discussion Papers on Development Policy*. Vol. 246. Bonn.

Askitas, Nikolaos, and Klaus F Zimmermann. 2009. “Google Econometrics and Unemployment Forecasting.” *Applied Economics Quarterly* 55 (2): 107–20. doi:10.3790/aeq.55.2.107.

———. 2015. “The Internet as a Data Source for Advancement in Social Sciences.” *International Journal of Manpower* 36 (1): 2–12. doi:10.1108/IJM-02-2015-0029.

Bańbura, Marta, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin. 2013. “Now-Casting and the Real-Time Data Flow.” In *Handbook of Economic Forecasting*, 2A:195–237. doi:10.1016/B978-0-444-53683-9.00004-9.

Bulut, Levent. 2017. “Google Trends and the Forecasting Performance of Exchange Rate Models.” *Journal of Forecasting*, no. September: 1–13. doi:10.1002/for.2500.

Carrière-Swallow, Yan, and Felipe Labbé. 2013. “Nowcasting with Google Trends in an Emerging Market.” *Journal of Forecasting* 32 (4): 289–98. doi:10.1002/for.1252.

Center for International Earth Science Information Network Columbia University. 2017. *Gridded Population of the World, Version 4, (GPWv4): Population Count, Revision 10*. 4.10. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). doi:10.7927/H4PG1PPM.

Choi, Hyunyoung, and Hal Varian. 2012. “Predicting the Present with Google Trends.” *Economic Record* 88 (SUPPL.1): 2–9. doi:10.1111/j.1475-4932.2012.00809.x.

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. “Detecting Influenza Epidemics Using Search Engine Query Data.” *Nature* 457 (7232). Nature Publishing Group: 1012–14. doi:10.1038/nature07634.

Google. 2018a. “Google Correlate.” https://www.google.com/trends/correlate.

———. 2018b. “Google Trends.” https://www.google.com/trends.

GSMA. 2018. “The Mobile Economy in Sub-Saharan Africa 2018.” London.

https://data.worldbank.org/region/sub-saharan-africa.

Guzmán, Giselle. 2011. "Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations." *Journal of Economic and Social Measurement* 36 (3): 119–67. doi:10.3233/JEM-2011-0342.

Hamid, Alain, and Moritz Heiden. 2015. "Forecasting Volatility with Empirical Similarity and Google Trends." *Journal of Economic Behavior and Organization* 117. Elsevier B.V.: 62–81. doi:10.1016/j.jebo.2015.06.005.

International Telecommunication Unit. 2018. "World Telecommunication/ICT Development Report and Database." https://www.itu.int.

Kalkuhl, Matthias, Joachim von Braun, and Maximo Torero. 2016. "Volatile and Extreme Food Prices, Food Security, and Policy: An Overview." In *Food Price Volatility and Its Implications for Food Security and Policy*, edited by Matthias Kalkuhl, Joachim von Braun, and Maximo Torero, 3–31. Cham: Springer International Publishing. doi:10.1007/978-3-319-28201-5.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6167): 1203–5. doi:10.1126/science.1248506.

Lindh, Thomas. 2011. *Long-Horizon Growth Forecasting and Demography*. *The Oxford Handbook of Economic Forecasting*. Oxford University Press. doi:10.1093/oxfordhb/9780195398649.013.0022.

Marthews, Alex, and Catherine Tucker. 2014. "Government Surveillance and Internet Search Behavior." *Available at SSRN* 66 (1): 176–78. doi:10.2139/ssrn.2412564.

Poushter, Jacob, Caldwell Bishop, and Hanyu Chwe. 2018. "Social Media Use Continues to Rise in Developing Countries but Plateaus Across Developed Ones." *Pew Research Center*. Vol. June.

Preis, Tobias, Helen Susannah Moat, and H Eugene Stanley. 2013. "Quantifying Trading Behavior in Financial Markets Using Google Trends." *Scientific Reports* 3: 1684. doi:10.1038/srep01684.

Scott, Steven L., and Hal R Varian. 2013. "Bayesian Variable Selection for Nowcasting Economic Time Series." *Economics of Digitization.*

Seabold, Skipper, and Andrea Coppola. 2015. "Nowcasting Prices Using Google Trends An Application to Central America." *World Bank Policy Research Working Paper*, no. 7398.

Stephens-Davidowitz, Seth. 2013. “Essay Using Google Data.” Harvard University.

———. 2017. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. Dey Street Books.

Stephens-Davidowitz, Seth, and Hal R. Varian. 2014. “A Hands-on Guide to Google Data.” Mountain View, CA.

Stock, James H., and Mark W. Watson. 2015. *Introduction to Econometrics*. 3rd Editio. Pearson Education.

Suhoy, Tanya. 2009. “Query Indices and a 2008 Downturn: Isreali Data.” 2009.06. Discussion Paper Series of the Bank of Isreal. Jerusalem.

The Economist. 2017. “The European Commission Levies a Huge Fine on Google,” July 1. https://www.economist.com/news/business/21724436-its-case-not-perfect-it-asks-right-questions-european-commission-levies-huge.

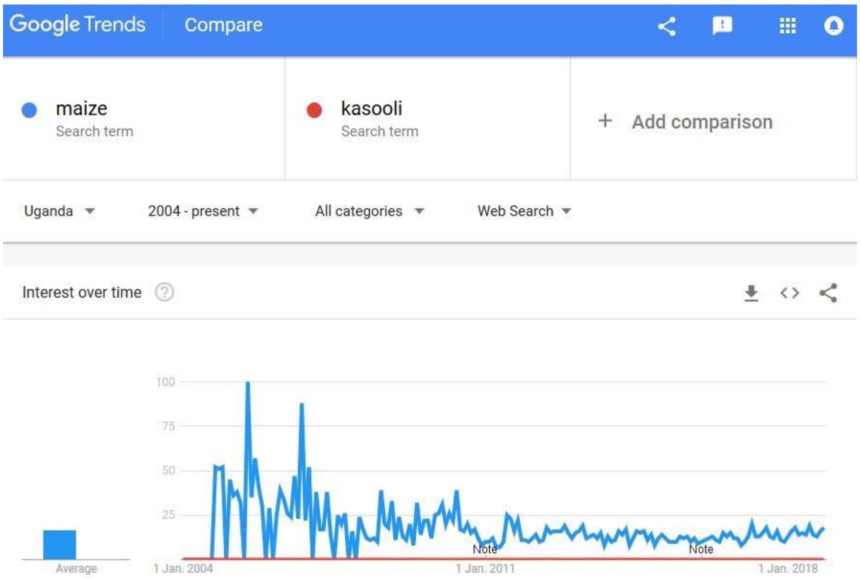
The World Bank. 2018. “World Development Indicators.” https://data.worldbank.org/indicator/IT.NET.USER.ZS.

Vosen, Simeon, and Torsten Schmidt. 2011. “Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends.” *Journal of Forecasting* 30 (6): 565–78. doi:10.1002/for.1213.

Weidmann, Nils B, S. Benitez-Baleato, Philipp Hunziker, Eduard Glatz, and Xenofontas Dimitropoulos. 2016. “Digital Discrimination: Political Bias in Internet Service Provision across Ethnic Groups.” *Science* 353 (6304): 1151–55. doi:10.1126/science.aaf5062.

World Bank Group. 2016. “World Development Report 2016: Digital Dividends.” Washington D.C.

Appendix



● maize
Search term

● kasooli
Search term

Uganda ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Figure A: Luganda vs. English: Search-Term Comparison for Uganda.

Source: Screenshot taken from https://trends.google.com on Nov, 11th 2018.

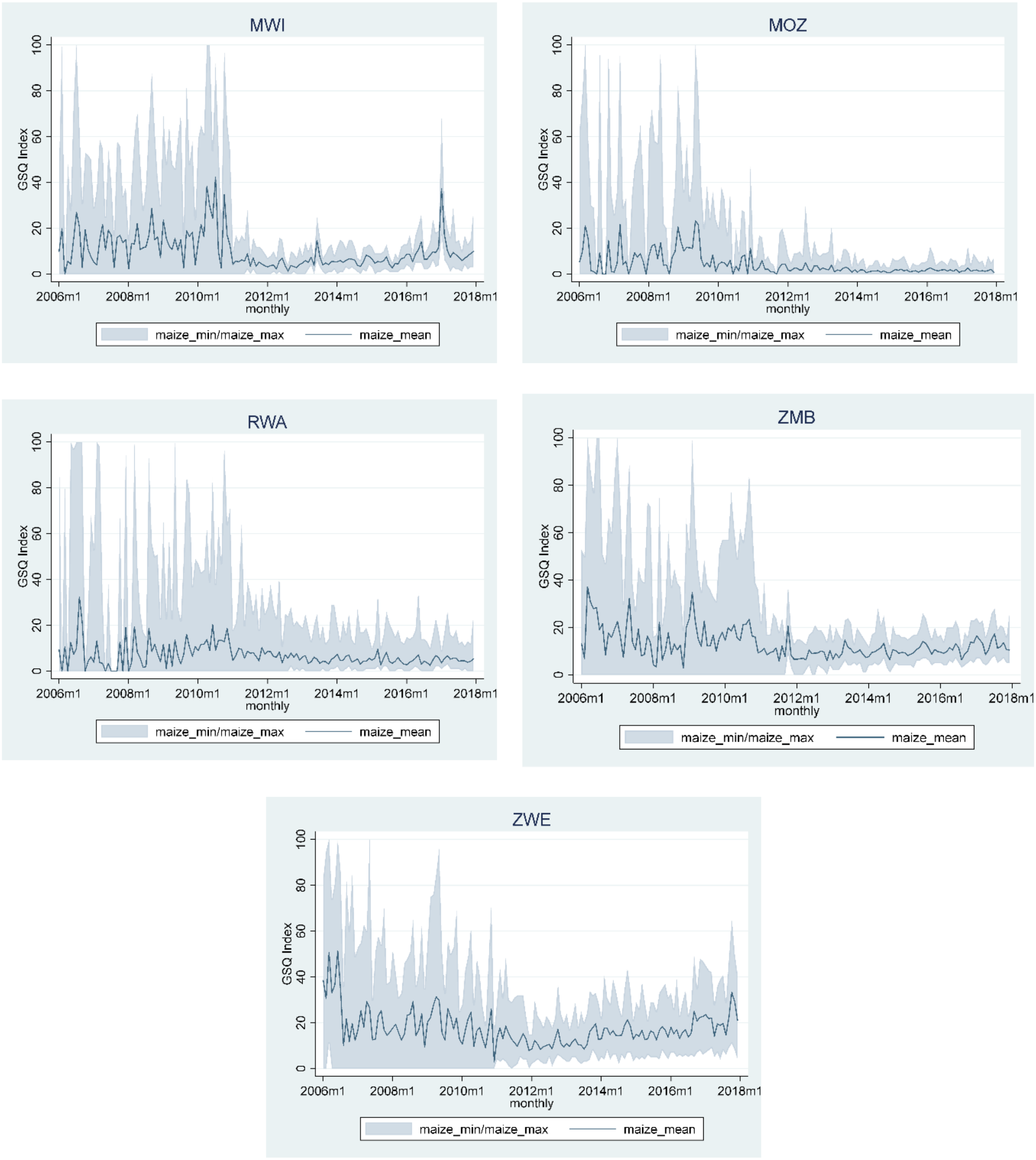


Figure B: Sampling Noise of GSQ Data for the Term *maize* in Malawi, Mozambique, Rwanda, Zambia and Zimbabwe.

Source: Own compilation based on data extracted from www.google.de/trends, sampled over a period of 30 days.

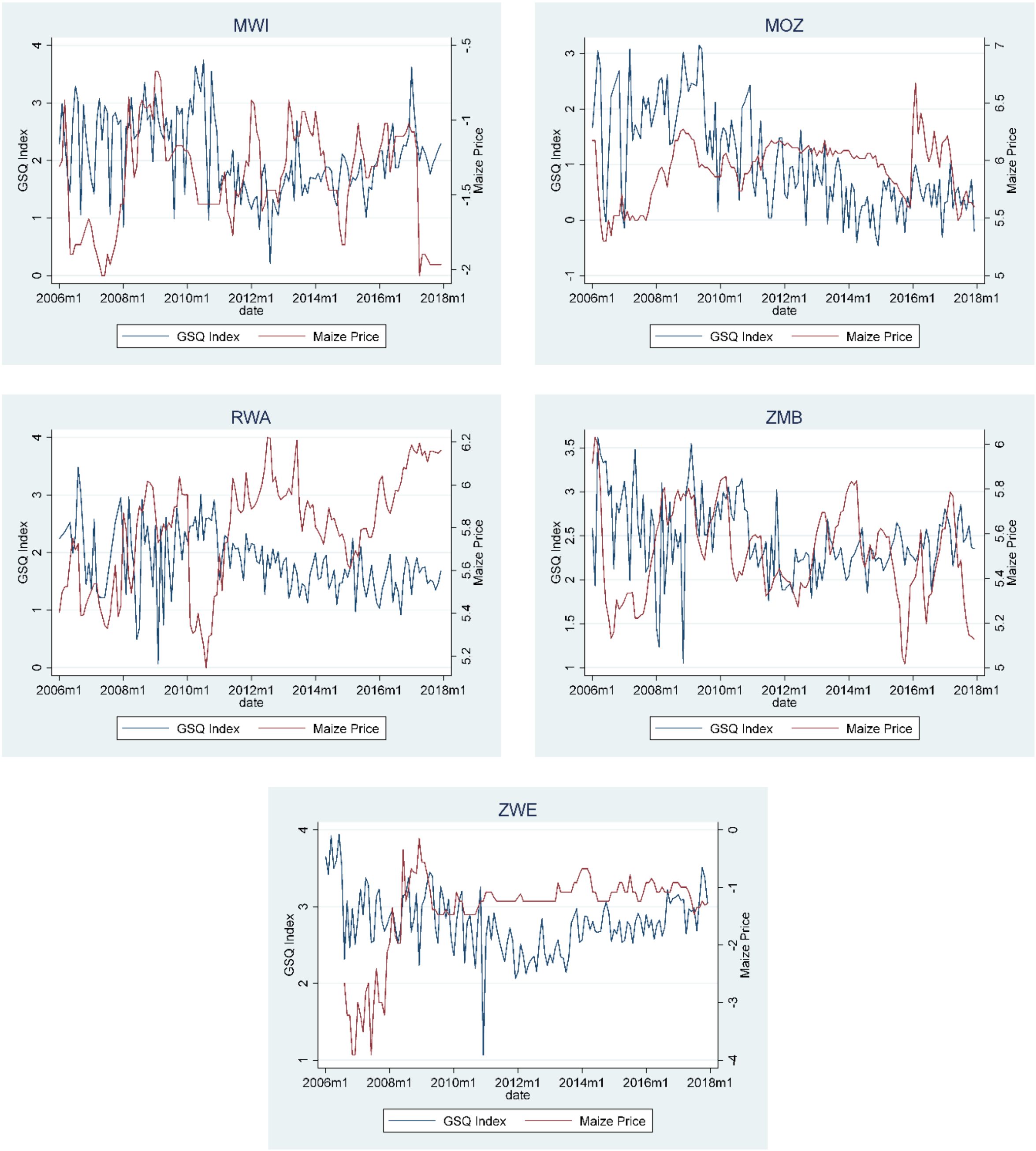


Figure C: Maize Prices and GSQ Data for the term *maize* in Malawi, Mozambique, Rwanda, Zambia and Zimbabwe.

Source: Own compilation.

	Maize Price		GSQ Volume	
Country	Mean	SD	Mean	SD
ETH	5.52	0.29	2.21	0.54
KEN	5.74	0.27	3.10	0.28
MOZ	5.93	0.25	1.03	0.91
MWI	-1.38	0.34	2.09	0.68
RWA	5.80	0.25	1.83	0.55
TZA	5.65	0.36	2.23	0.44
UGA	5.46	0.33	2.69	0.33
ZMB	5.51	0.21	2.46	0.43
ZWE	-1.39	0.72	2.79	0.40

Table A: Summary Statistics of the Logged Price and GSQ Series.

Source: Own Compilation.

Country	Variable	P-Perron Statistic	P-Perron Lags	Order of Integration
ETH	maize_ln	-4.904684	4	I(0)
		0.00		
	maize_usd_ln	-2.704097	4	I(0)
		-0.07		
KEN	maize_ln	-6.431211	4	I(0)
		0.00		
	maize_usd_d1	-9.468563	4	I(1)
		0.00		
MOZ	maize_ln	-4.860092	4	I(0)
		0.00		
	maize_usd_ln	-3.165808	4	I(0)
		0.02		
MWI	maize_ln	-7.32375	4	I(0)
		0.00		
	maize_usd_ln	-2.916184	4	I(0)
		0.04		
RWA	maize_ln	-8.677816	4	I(0)
		0.00		
	maize_usd_ln	-2.604109	4	I(0)
		0.09		
TZA	maize_ln	-6.34058	4	I(0)
		0.00		
	maize_usd_ln	-2.52525	4	I(0)
		0.11		
UGA	maize_ln	-8.609276	4	I(0)
		0.00		
	maize_usd_ln	-3.169517	4	I(0)
		0.02		
ZMB	maize_ln	-8.130215	4	I(0)
		0.00		
	maize_usd_ln	-3.444542	4	I(0)
		0.01		
ZWE	maize_ln	-7.265877	4	I(0)
		0.00		
	maize_usd_d1	-16.5487	4	I(1)
		0.00		

Table B: Philipps-Perron Unit Root Test Statistic.

Note: maize ln = GSQ search term *maize*, maize usd ln = local maize prices in USD, d1 = first differences, p-values in parentheses. Source: Own estimation.

Country	lag	LL	LR	p-value	SBIC
ETH	0	-22.230			0.36
	1	120.284	285.03	0.00	-1.70
	2	127.406	14.24	0.00	-1.77
KEN	0	133.677			-1.90
	1	136.566	5.78	0.02	-1.91
MOZ	0	2.550			0.00
	1	108.051	211.00	0.00	-1.49
MWI	0	-45.007			0.69
	1	65.392	220.80	0.00	-0.88
RWA	0	-5.470			0.11
	1	111.394	233.73	0.00	-1.54
TZA	0	-55.493			0.84
	1	86.046	283.08	0.00	-1.18
UGA	0	-46.221			0.71
	1	55.093	202.63	0.00	-0.73
ZMB	0	29.229			-0.39
	1	153.266	248.07	0.00	-2.15
	2	164.714	22.90	0.00	-2.28
ZWE	0	-29.159			0.48
	1	-23.402	11.52	0.00	0.43
	2	-16.778	13.25	0.00	0.37

Table C: Lag-Order Selection Statistics.

Note: LL=log likelihood, LR=likelihood ratio, SBIC=Schwarz's Bayesian Information Criterion, maximum number of 6 lags included. Source: Own estimation.

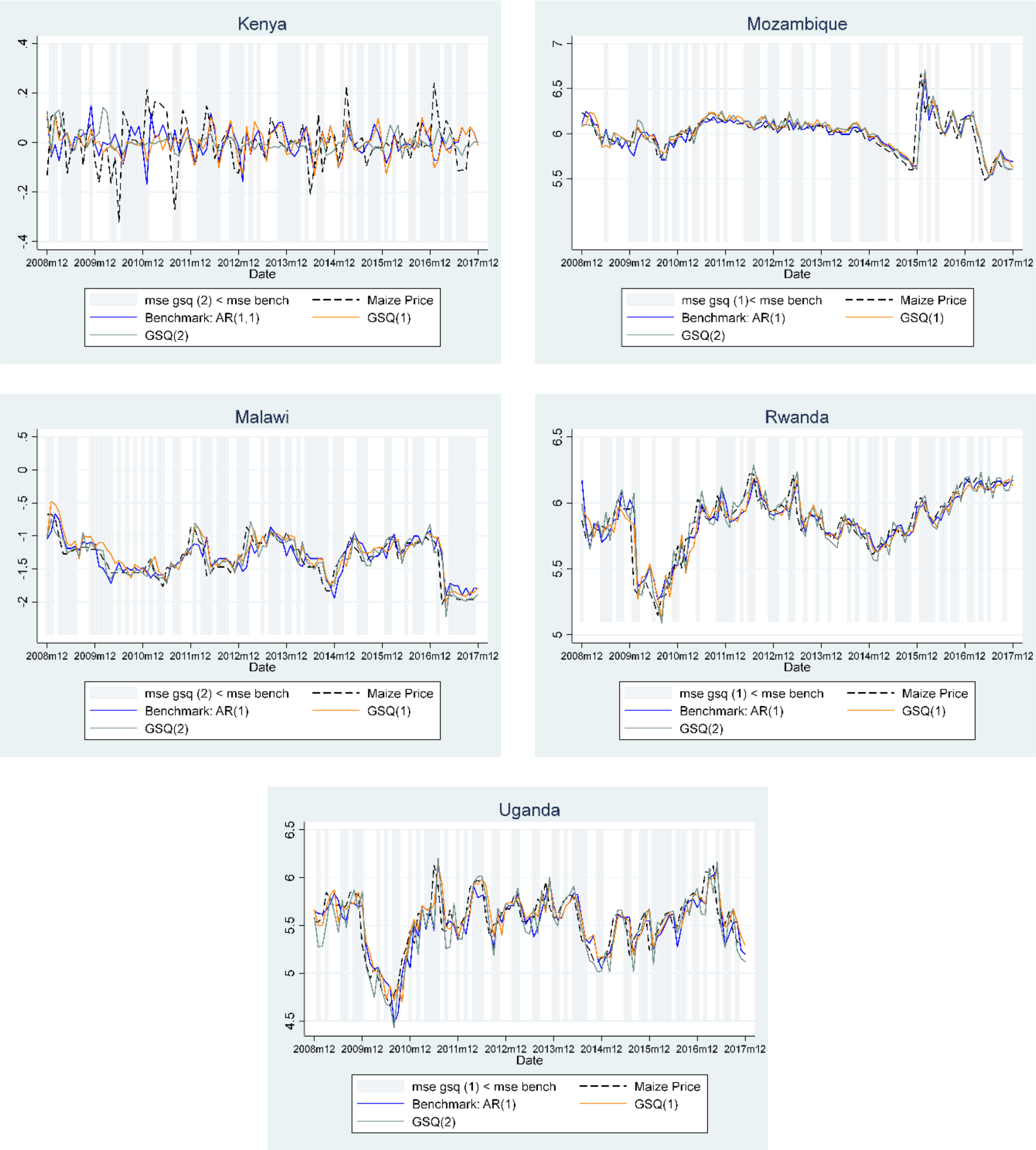


Figure D: Benchmark vs. GSQ-Augmented Out-Of-Sample Forecasts.

Note: In-sample training period (01.2006 - 12.2018) not displayed. Source: Own estimation.