Organisation for Economic Co-operation and Development

DSTI/CDEP/MADE(2023)13/FINAL

Unclassified

English - Or. English

7 June 2024

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INNOVATION
COMMITTEE ON DIGITAL ECONOMY POLICY

Working Party on Measurement and Analysis of the Digital Economy

Nowcasting the growth rate of the ICT sector

JT03545501

Foreword

This paper details the methodology used to nowcast the growth rate of the information and communication technology (ICT) sector in "Chapter 1: The growth outlook of the ICT sector" of the *OECD Digital Economy Outlook 2024 (Volume 1)*. It was written by Camilo Umana Dajud with contributions from Nicolas Benoit. This paper was approved and declassified by the OECD Digital Policy Committee on 4 December 2023 and prepared for publication by the OECD Secretariat.

Note to Delegations:

This document is also available on iLibrary as Umana Dajud, C. (2024), "Nowcasting the growth rate of the ICT sector", OECD Digital Economy Papers, No. 362, OECD Publishing, Paris, https://doi.org/10.1787/eb4938a0-en.

Table of contents

FIGURES

TABLES

1 Background

In an era characterised by rapid digital transformation, leveraging innovative data sources for economic measurement has gained considerable attention. Among these sources, Google Trends data has emerged as a promising tool for tracking economic activity in real time. This paper delves into the nuanced landscape of utilising Google Trends data to nowcast growth rates. Despite the growing interest in this domain, the literature lacks a consensus on the most effective methodologies, leading to a diverse array of statistical approaches.

In this respect, this technical paper explains in detail the methodological choices used to nowcast ICT sector growth using Google Trends data. Rather than relying on a single economic theory or statistical method, the paper adopts a purely empirical approach, focusing on the machine learning methods and statistical techniques to extract relevant information from Google Trends data that yield the most accurate results. The methodology encompasses several key stages, including data extraction from Google Trends, filtering out noise in the data brought by downward trends, sampling noise and seasonality, and employing a tailored combination of statistical and machine learning techniques to make growth rate predictions.

By adopting a data-driven methodology and addressing data challenges, this paper aims to contribute to the advancement of economic measurement of the digitalisation of the economy, offering insights into the growth dynamics of the ICT sector and potentially informing policy decisions for harnessing the benefits of the ongoing digital transformation. This technical paper complements Chapter 1 of the O*ECD Digital Economy Outlook 2024 (Volume 1)* (OECD, 2024[1]).

The paper is structured as follows: first, it reviews the existing literature on using Google Trends data for the measurement of economic activity; then it outlines the process of extracting relevant insights from Google Trends data, highlighting challenges and correction strategies; next, it details the computation of ICT sector growth rates using OECD's Structural Analysis (STAN) database; then, it offers a comprehensive analysis of the selection process for statistical methods and machine learning models, focusing on performance metrics; and finally, it presents the results before concluding.

2 Measuring economic activity with Google Trends data

Since the seminal paper by Choi and Varian (2012[2]), there has been a growing interest in using Google Trends data to measure economic activity. The literature focuses on two main areas: nowcasting and forecasting. Nowcasting refers to the use of Google Trends data to estimate current economic activity. Forecasting refers to the use of Google Trends data to predict future economic activity.

At the same time, amidst the vast influx of novel data stemming from digital transformation, there is still a lack of data on many of its facets. Foremost, there is a lack of up-to-date cross-country data on the economic performance of the sector at the core of the digital transformation: the ICT sector. This lack of up-to-date data on the economic performance of the ICT sector limits the evaluation and design of public policies to unleash the benefits of digital transformation.

In the case of sectoral growth rates, the lack of up-to-date data is especially severe. The OECD STAN database is the main source of data on the economic performance of the ICT sector (Horvát and Webb, 2020[3]). However, STAN data is only available until 2018 for most countries and for a few countries until 2019. As a result, cross-country data on the economic performance of the ICT sector is only available with a lag of three to four years. It is precisely this gap that the nowcasting model aims to fill.

Google Trends data has been used in a number of occasions to nowcast economic activity. In Choi and Varian's seminal paper Google Trends data is used to nowcast automobile sales, home sales, retail sales, and travel behaviour among other economic time series. Later OECD's Weekly Tracker of economic activity (OECD, n.d.[4]) brought together Choi and Varian's (2012[2]) insights with machine learning methods to nowcast weekly GDP growth. The tracker provides weekly GDP figures for 46 countries starting in 2020 (OECD, n.d.[4]).

Other efforts to measure economic activity in real time include nowcasting GDP growth in Brazil (Bantis et al., 2023[5]), the United States (Bantis et al., 2023[5]); (Kohns, David and Bhattacharjee, 2023[6]), Finland (Heikkinen and Joni, 2019[7]) and Germany (Götz and Knetsch, 2019[8]). In the case of Germany, Götz and Knetsch (2019[8]) use Google Trends data as the input of a bridge equation model to nowcast aggregate GDP, various GDP components, as well as monthly activity indicators. For the US, Bantis et al., (2023[5]) use a dynamic factor model to measure US GDP in real time. Similarly, Kohns, David and Bhattacharjee (2023[6]) mixed frequency Bayesian Structural time series model to the same aim.

The modelling approach in this technical paper is directly inspired by OECD's Weekly Tracker methodology (OECD, n.d.[4]). It also builds on other efforts to nowcast economic activity using Google Trends data. In this regard, this paper uses a purely empirical theory agnostic approach.

3 Extracting information from Google Trends data

As demonstrated by previous efforts to measure economic activity in real time, Google Trends data contains relevant information on economic performance. This section describes Google Trends data and how it is processed to extract useful information for nowcasting the growth of the ICT sector.

Description of Google Trends data

Google Trends does not provide the exact number of searches for a given keyword. Instead, Google Trends provides a relative measure of the number of searches for a given keyword that is computed as follows. First, the platform counts the number of searches for a given keyword within a particular time frame and region. Second, it determines the aggregate number of searches for all keywords in the same time frame and region. Third, a ratio is computed by dividing the keyword-specific search count by the total search count. Lastly, this ratio is multiplied by 100 to derive a relative measurement for the keyword's search volume (referred to in this technical paper as search indexes).

Real time data vs non-real time data

Google Trends provides two different types of data: real time data and non-real time data. Real time data is available for the last seven days and is updated every hour. Non-real-time data is available from January 2004 to the present and is updated daily. The data is available at the country level and at the regional level.

The growth rates of the ICT sector are computed using data that is only available on a yearly basis. For this reason, the nowcasting methodology described in this document is based on Google Trends non-real time data. An additional advantage of using non-real time data is that, unlike Google Trends real time data, is does not suffer from frequency inconsistency, as noted by Eichenauer, Indergand and Martínez (2022[9]). Frequency inconsistency refers here to the lack of consistency between daily and monthly Google Trends time series.

Sampling noise

Google Trends non-real time data suffer from sampling noise. To protect privacy, Google does not use all searches to compute search indexes. Instead, search indexes are computed using a random sample of all searches. The size of the random sample is not disclosed by Google. Taking random samples of all searches introduces variability each time search indexes are fetched. This variability is known as sampling noise. The sample noise from Google Trends search indexes is more severe for smaller regions and/or less popular search categories (Eichenauer, Indergand and Martínez, 2022[9]). For those regions or search categories the universe of searches is smaller. As a result, the random sample is more likely to be unrepresentative.

To reduce sampling noise and increase the consistency of search indexes two corrections are implemented: the use of multiple samples and variance correction.

Multiple samples

The first correction to reduce sampling noise is to fetch from Google Trends five different samples for every region and category. The average obtained from these samples is the search index used in the nowcasting of ICT growth rates. Previous research has shown that taking several samples from the Google Trends API reduces sampling noise up to 90% (Eichenauer, Indergand and Martínez, 2022[9]).

Table 3.1 reports the variance of five distinct samples alongside the variance of their respective averages. Notably, the table shows that the variance of the average of the five samples is 658.90. This is almost 4% lower than the variance of the five individual samples. This observation implies that fetching multiple samples from the Google Trends API effectively mitigates sampling noise.

Table 3.1. Comparing sample variance and the variance of sample averages

Sample	Variance
Sample #1	684.32
Sample #2	683.84
Sample #3	683.49
Sample #4	683.15
Sample #5	682.17
Mean of Samples	658.90

Note: The table shows the variance of five different samples and the variance of the average of these five samples.
Source: Author's calculations based on Google Trends data.

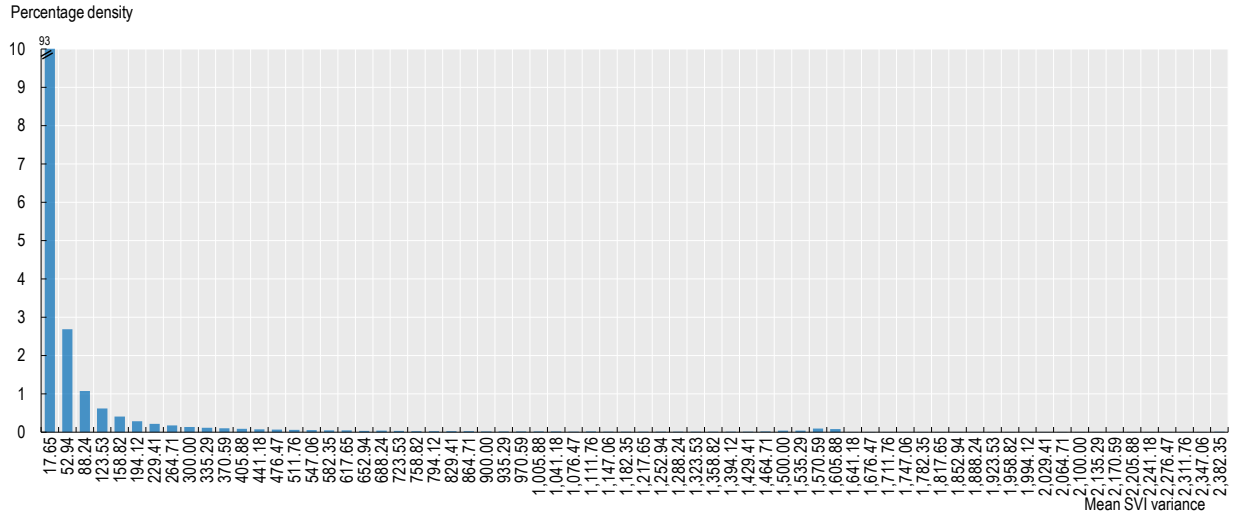
Variance correction

The second strategy to reduce sampling noise is to drop search indexes with a variance exceeding 10 across the five different samples. This ensures that search indexes that are computed by Google Trends using a sample too small to be representative are dropped and not used in the nowcasting model.

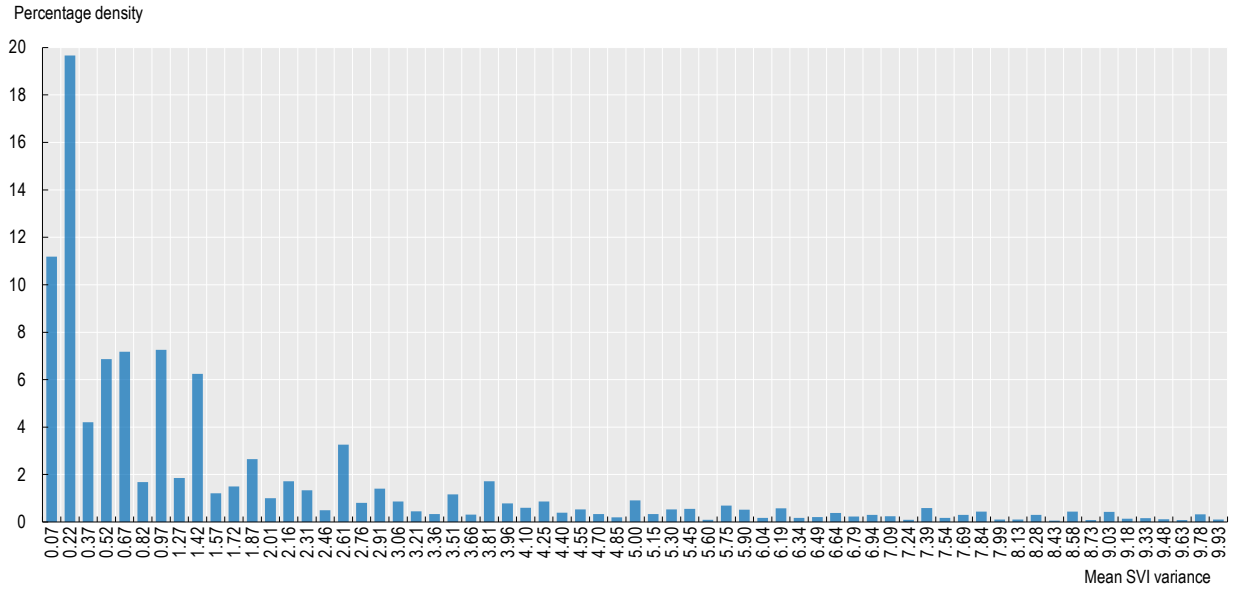
Panels a) and b) in Figure 3.1 show the distribution of variance in the average search index for the full sample, as well as after excluding search indexes with a variance surpassing 10. Notably, the figures show that a substantial share of search indexes is characterised by remarkably high variances. Specifically, the data reveals that nearly 20% of search indexes exhibit variances exceeding 10 – a significant percentage considering the mean across search indexes stands at 25.6.

Figure 3.1. Sampling noise distribution before and after correction

a. SVI variance distribution – Full sample



b. SVI variance distribution – Restricted sample



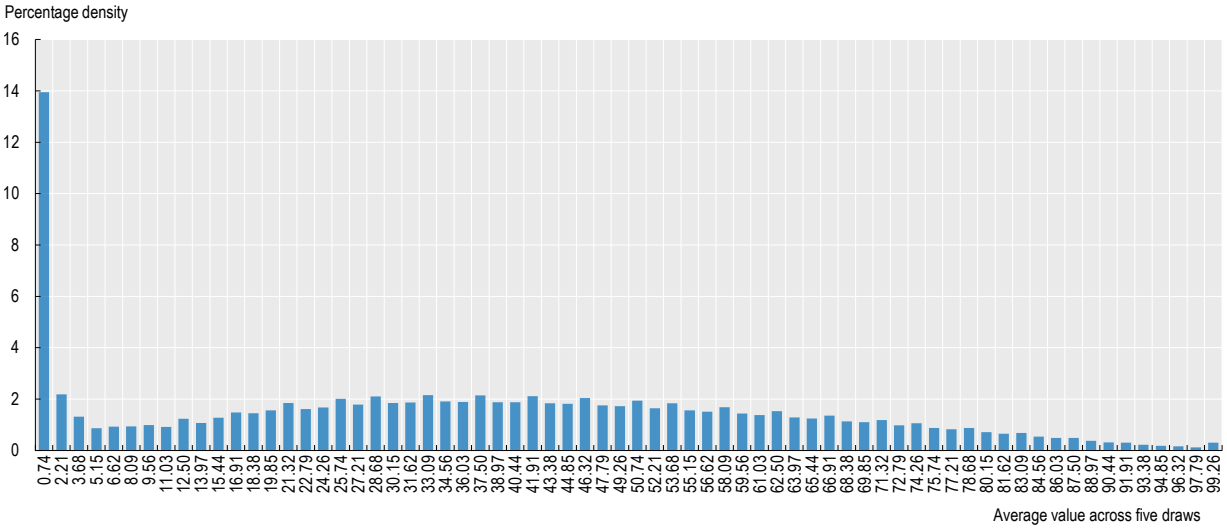
Notes: Panel a shows the distribution of the variance of the average search index for the full sample. The average search index is computed using five different samples from Google Trends API. Panel b shows the distribution of the variance of the average search index after dropping search indexes with a variance exceeding 10 across five draws. The average search index is computed for every region and category using five different samples from Google Trends API.
Source: Author's calculations based on Google Trends data.

Figure 3.2 panel a presents the distribution of the average search index, calculated from the five distinct samples irrespective of their variances. Panel b in Figure 3.2 shows the distribution of the average search index is depicted after removing search indexes with variances surpassing 10 across these five samples. Both figures indicate that eliminating search indexes with variances exceeding 10 results in a moderate reduction in the average search index's variance. Importantly, these visuals highlight that such exclusion

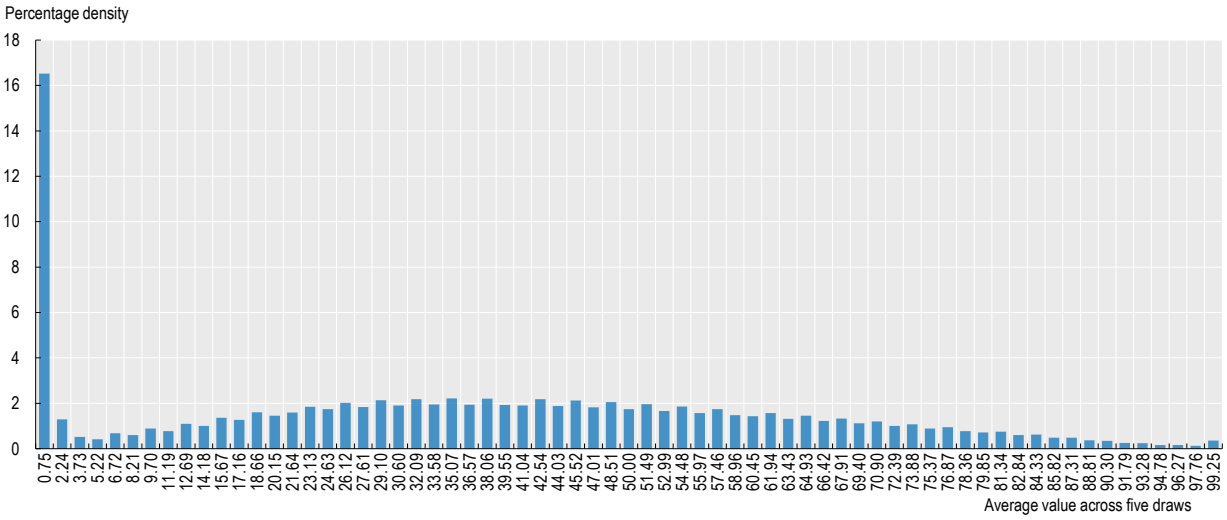
has no impact on the distribution of the average search index. This observation implies that the exclusion of search indexes with variances exceeding 10 does not introduce any bias into the nowcasting model.

Figure 3.2. Average search index distribution before and after correction

a. Average search index distribution – Full sample



b. Average search index distribution – Restricted sample



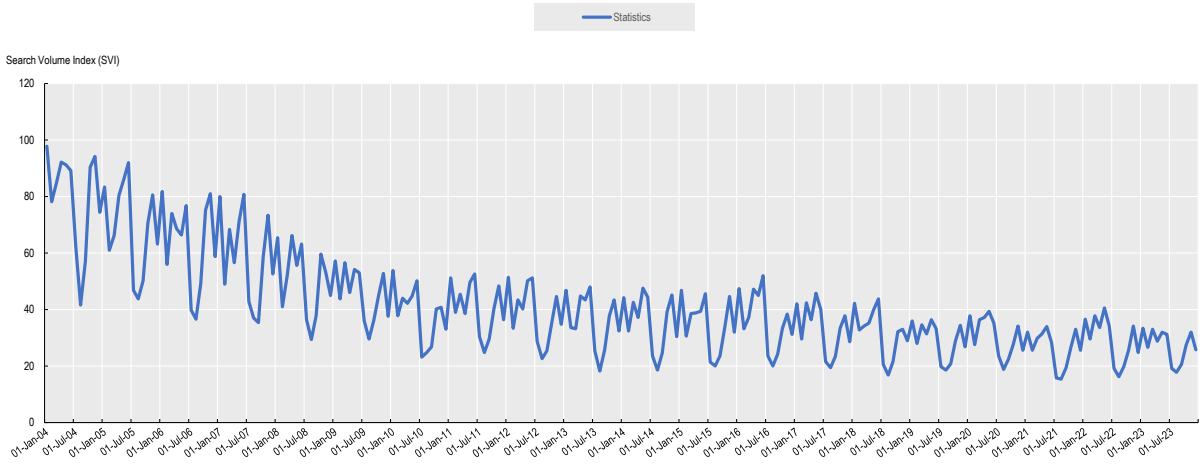
Notes: Panel a shows the distribution of the average search index for the full sample. The average search index is computed for every region and category. Panel b shows the distribution of the average search index after dropping search indexes with a variance exceeding 10 across five draws. The average search index is computed for every region and category using five different samples from Google Trends API.
Source: Author's calculations based on Google Trends data.

In summary, the nowcasting model uses the average of five samples and drops search indexes with a variance exceeding 10 across the five samples. This correction reduces sampling noise and increases the consistency of search indexes.

Downward trend

Another issue with Google Trends data is that many search indexes trend downward. The downward trend does not mean, however, that interest in each topic has decreased overtime. Instead, the downward trend is because the number of searches for all keywords has increased during this period. As a result, the relative number of searches for a given keyword decreases overtime. Figure 3.3 shows the search index for the category "Statistics" in Austria. The figure exhibits a downward trend. However, the downward trend should not be interpreted as a decrease in interest in statistics, but rather because of the overall increase in the number of searches for all keywords.

Figure 3.3. Search index for the category "statistics" in Austria



Source: Author's calculations based on Google Trends data.

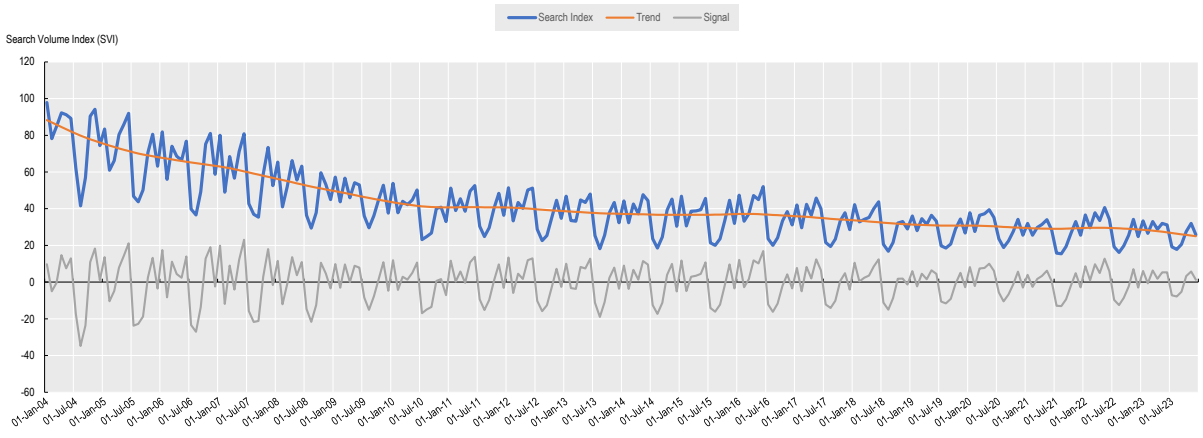
Note, however, that the downward trend is more pronounced during the first part of the period. This is since Google Trends data is normalised using the number of searches for all keywords. While the number of searches for all keywords has increased over time, this increase was notably more rapid during the initial years of Google Trends data collection. As a result, the downward trend is more pronounced during the first years of Google Trends data.

Various statistical methods exist for mitigating the impact of downward trends in search indexes. While developing the nowcasting model employed in this technical paper, multiple alternatives were explored and tested.

Hodrick-Prescott filter

A first option is to apply a Hodrick-Prescott filter. The Hodrick-Prescott filter is a statistical method that separates a time series into a trend component and a cyclical component. It is a widely used method to filter out the downward trend in search indexes (Eichenauer, Indergand and Martínez, 2022[9]). Figure 3.4 shows the results of applying this filter to the search index for the "statistics" category in Austria. This figure shows that the Hodrick-Prescott filter can filter out the downward trend in search indexes. The cyclical component of the Hodrick-Prescott filter could potentially be used in the nowcasting model.

Figure 3.4. Search index for the "statistics" category in Austria decomposed using a Hodrick-Prescott filter



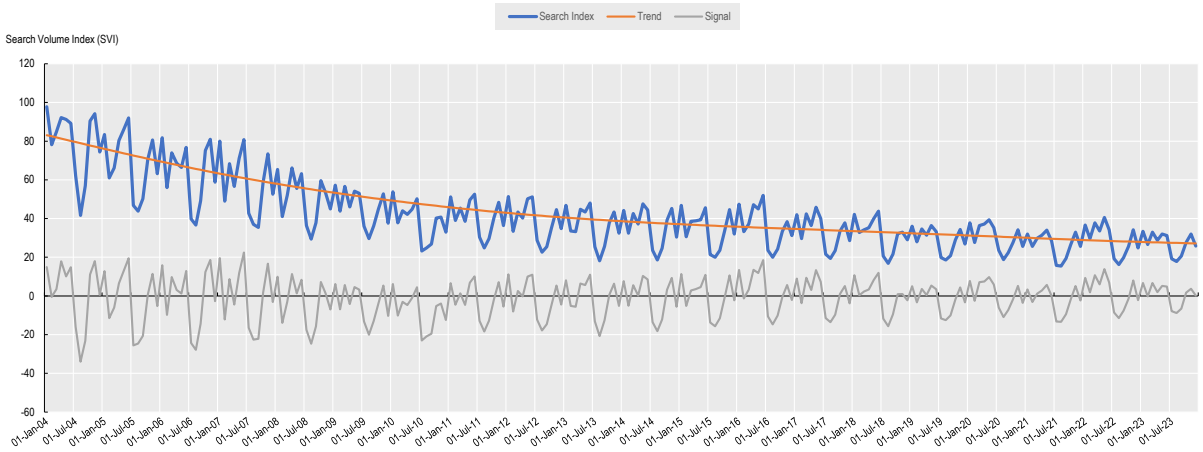
Source: Author's calculations based on Google Trends data.

Lowess smoothers

An alternative to correct for the downward trend is possible by locally weighted scatterplot smoothing (hereafter Lowess smoothers). Lowess smoothers are the result of a non-parametric regression method that fits a smooth curve to a scatterplot. In doing so, the smoother is able to filter out the downward trend in search indexes.

Figure 3.5 shows the result of applying a Lowess smoother to the search index for the "statistics" category in Austria. It shows that the Lowess smoother is able to filter out the downward trend in search indexes and therefore could potentially be used in the nowcasting model.

Figure 3.5. Search index for the statistics category in Austria decomposed using a Lowess smoother



Source: Author's calculations based on Google Trends data.

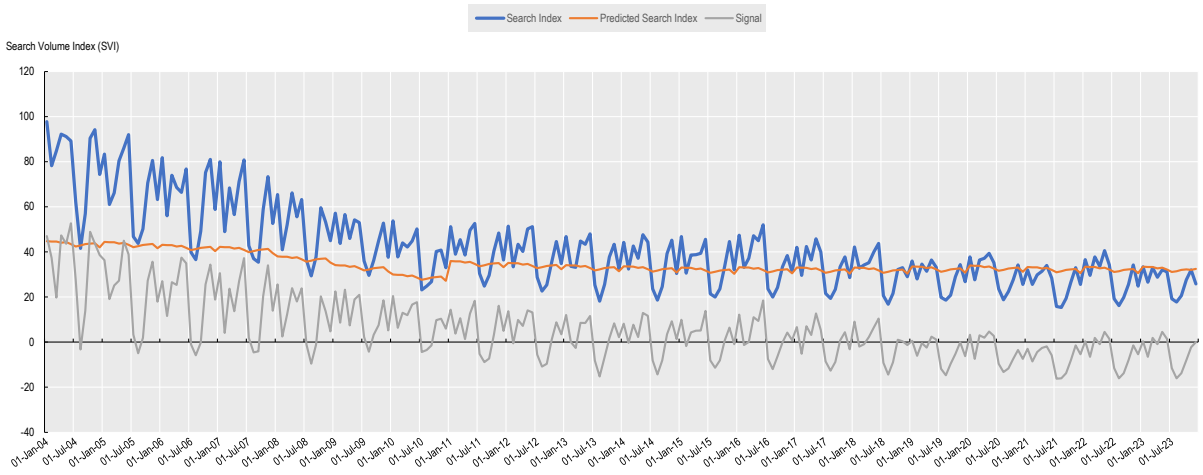
Fixed effects

A third and final alternative to mitigate the impact of downward trends in search indexes is to use fixed effects. Fixed effects are a statistical method that allows to filter out common components in panel data. In the case of Google Trends, the search indexes vary along time and country dimensions. Fixed effects allow thus to filter out the common component along time and/or country dimensions.

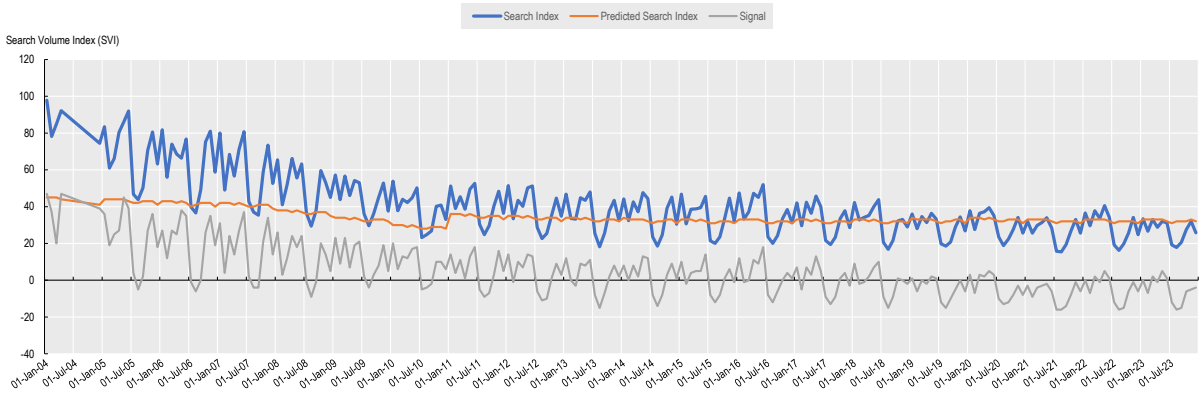
Figure 3.6 panel a shows the result of applying country, month, and year fixed effects to the search index for the statistics category in Austria. In this case, this is done by running an ordinary least square (OLS) regression (Figure 3.6, panel a). Alternatively, a Poisson regression could be used. Figure 3.6, panel b shows the results of doing so.

Figure 3.6. Search index for the "statistics" category in Austria decomposed using fixed effects

a. Fixed effects plotted with an OLS regression



b. Fixed effects plotted with a Poisson regression



Source: Author's calculations based on Google Trends data.

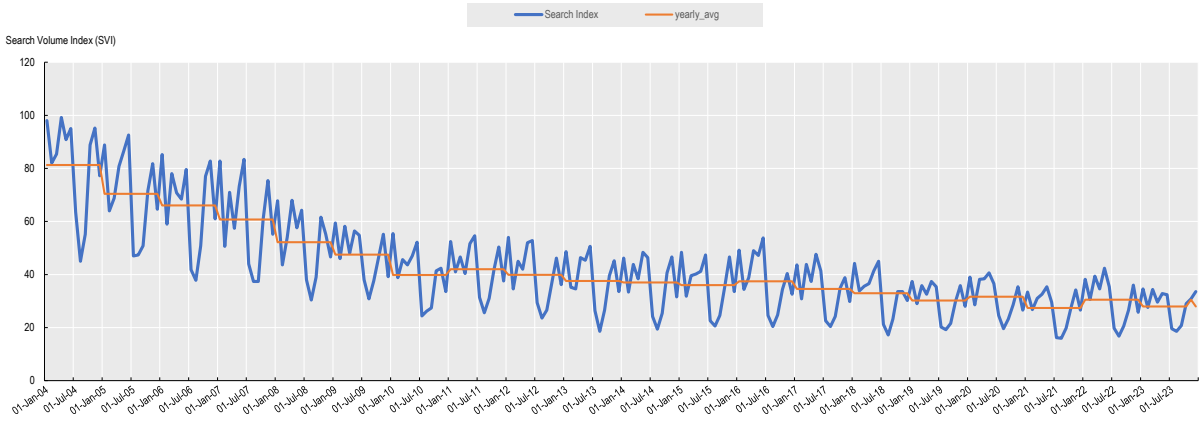
Comparing different methods

This paper follows a purely empirical and theory agnostic approach. While visually the Hodrick-Prescott filter and Lowess smoother seem to be able to filter out the downward trend in search indexes, it is not clear which method is the most appropriate. To select the most appropriate method, the performance of each method is compared using the nowcasting model described below as well as the root mean squared error (RMSE).

Seasonality

An additional limitation of Google Trends data is that it is affected by seasonality, as demonstrated in Figure 3.7, as well as in Figure 3.3 and Figure 3.4. The seasonal pattern arises because the number of searches for a given keyword is not constant across time. Instead, the number of searches for a given keyword is higher during certain months of the year. For example, the number of searches for the keyword "ski" is higher during the winter months. The seasonal pattern is more pronounced for some keywords than for others. For example, the seasonal pattern is more pronounced for the keyword "ski" than for the keyword "statistics".

Figure 3.7. Search index for the "statistics" category in Austria: Monthly versus yearly average



Source: Author's calculations based on Google Trends data.

The seasonal fluctuations in search indexes could potentially pose challenges for the nowcasting model. However, as elaborated just after, the data required for calculating the growth rates of ICT sector is only available on an annual basis. Consequently, the model bypasses the use of monthly data in favour of yearly averages. demonstrates that utilising yearly averages effectively neutralises the impact of seasonal trends in search indexes. Furthermore, the near-zero correlation between the yearly averaged search indexes and the annual growth rates of the ICT sector implies that the model is unlikely to be influenced by the seasonality inherent in Google Trends data. To ensure robustness, this paper later details how the nowcasting model incorporates seasonal dummy variables as an additional safeguard against any lingering seasonal patterns.

4 Computing ICT growth rates

Nowcasting ICT growth rates using Google Trends requires two different types of data: ICT growth rates and Google Trends search indexes. While aggregate GDP growth rates are readily available, sectoral growth rates that are comparable across economies are not. This section describes how ICT growth rates are computed using OECD STAN data.

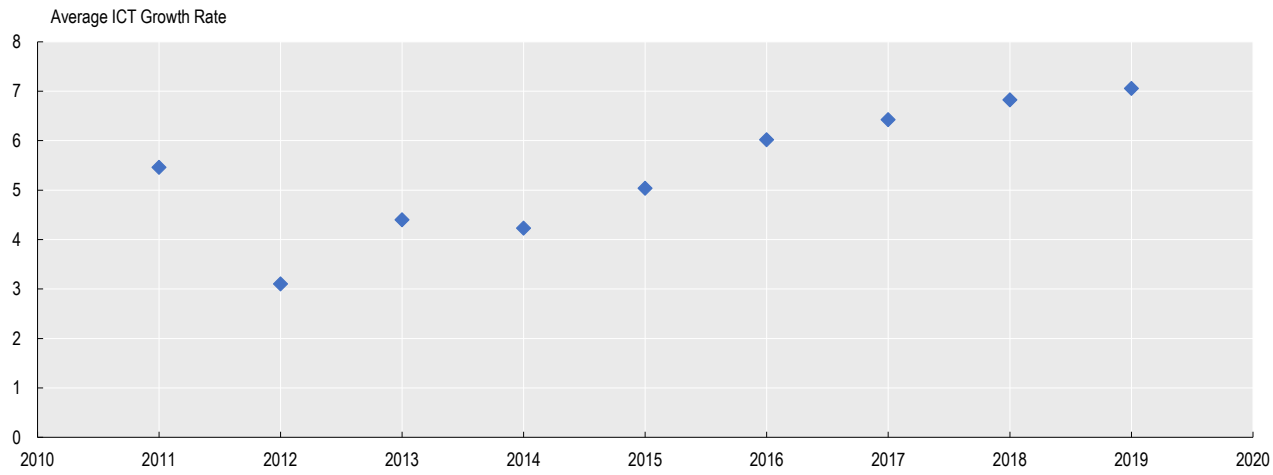
To compute the growth rates of the ICT sector, this paper uses the OECD STAN database. The database provides annual growth rates of value added by sector and is available for 35 OECD countries and 26 partner economies. STAN is currently available from 1970 to 2018, or 2019 for some countries.

An important characteristic of STAN is that it follows a standard industry list (Horvát and Webb, 2020[3]). This list allows for direct comparisons across countries. STAN industry list is based on the International Standard Industrial Classification of All Economic Activities (ISIC). However, to allow for longer time series, STAN industry list includes non-standard aggregates. Among these, STAN includes a category for the ICT sector at the two-digit level. The ICT sector is defined as the aggregate of the following ISIC divisions: 26, 61 and 62-63.

ICT sector growth rates

STAN provides annual data on gross value added by sector, along with the necessary deflators to adjust for inflation. These STAN figures serve as the basis for calculating the growth rates of the ICT sector. As illustrated in Figure 4.1, the average growth rates for the ICT sector in OECD countries have consistently been positive for each year encompassed by the STAN dataset. Notably, except for two years, the sector's growth rates show an upward trajectory throughout the period covered.

Figure 4.1. Average observed ICT growth rate by year in OECD countries



Source: Author's calculations using OECD STAN database.

The limitations of the OECD STAN database

Highlighting the challenges associated with securing comparable sectoral data across economies, the STAN dataset presents two significant limitations for this study. The first limitation lies in the temporal scope of the data: STAN's information is mostly current only up to 2018 for the majority of countries and extends to 2019 for a few economies. Addressing this temporal gap is precisely the primary objective of the nowcasting model. The second limitation pertains to the database's geographical coverage, specifically concerning the ICT sector aggregate. Although STAN includes data for 35 OECD countries and partner economies, the ICT sector aggregate is not universally available for all OECD members. Specifically, this aggregate is missing for the following OECD countries: Australia, Chile, Colombia, Costa Rica, Ireland, Israel, Japan, Korea, Luxembourg, New Zealand, and Republic of Türkiye (hereafter "Türkiye").

5 Choosing the best model

Statistical method selection

As explained previously, different statistical methods can be used to filter out the downward trend in search indexes. Following a purely empirical theory agnostic approach, the nowcasting model uses the statistical method that performs best. Table 5.1 shows the performance of each statistical method. The performance of each statistical method is compared using the RMSE of the validation data.

The RMSE of the validation data is computed as follows. First, the nowcasting model is estimated using the training data. Second, the nowcasting model is used to predict the growth rates of the ICT sector for the validation data. Third, the RMSE of the validation data is computed as the root mean squared error between predicted and observed ICT sector growth rates. The training data is the data from 2004 to 2017. The validation data is the data from 2018 to 2019.This approach ensures that the most effective statistical method is adopted for a more accurate and reliable nowcasting model.

As indicated previously, various statistical methods can be used to eliminate the downward trend observed in search indexes. Following a purely empirical theory agnostic approach, the nowcasting model incorporates the statistical technique that exhibits the highest performance. Table 5.1 details the RMSE of the validation data, which is the primary criterion for comparison.

Table 5.1. Comparing different statistical methods

	Hodrick-Prescott Levels	Lowess Levels	Fixed Effects	Hodrick-Prescott Log	Lowess Log	Hodrick-Prescott + Lowess
RMSE training data	0.32	2.55	2.67	0.0001	0.05	2.53
RMSE validation data	2.71	2.92	2.86	3.41	2.89	2.58

Note: The table shows RMSE for the training and validation data for different statistical methods.
Source: Author's calculations using OECD STAN database and Google Trends data.

Table 5.1 shows that the best performance is obtained with the combination of the Hodrick-Prescott filter and a Lowess smoother. In contrast, when used individually, the Hodrick-Prescott and Lowess filters yield suboptimal results, while the fixed effects method lags even further behind. The superior performance of the combined Hodrick-Prescott and Lowess smoother can likely be attributed to their complementary capabilities: the Hodrick-Prescott filter effectively eliminates the downward trend in search indexes, while the Lowess smoother partially filters out seasonal fluctuations in search indexes.

Additionally, the stand-alone use of either the Hodrick-Prescott filter or the Lowess smoother, as well as the fixed effects method, leads to overfitting of the training data, resulting in poor performance on the validation set. In stark contrast, the combined approach of using both the Hodrick-Prescott filter and the Lowess smoother avoids this pitfall of overfitting, thereby ensuring robust performance on the validation

data. As a result, the nowcasting model employs a two-stage approach, first applying the Hodrick Prescott filter to the search indexes, followed by the Lowess smoother.

Machine learning method selection

Another modelling choice concerns the machine learning method. This section compares the performance of different machine learning methods. The performance of each machine learning method is compared using the RMSE for the training and validation datasets. To choose the best performing model, several statistical alternative models were tested. This section reports only the three best performing models. The performance of a simple naive autoregressive model is also reported for comparison purposes.

Table 5.2 compares the performance the three different machine learning models using the root RMSE of the validation data. The table shows that the neural network has the lowest RMSE, and therefore the nowcasting model uses an artificial neural network. This type of network is based on a simplified probabilistic modelling of organic neurons. More precisely, the model is a multilayer perceptron. In this type of model, the connection between layers is established only in one direction avoiding any loops in the network.

Table 5.2. Comparison of different machine learning methods

	Two Stages	Gradient Boosting	AR(1)	Neural Network
RMSE training data	3.60	1.80	4.33	2.53
RMSE validation data	2.73	5.25	4.27	2.58

Note: The table shows the RMSE for the training and validation data for different machine learning methods.
Source: Author's calculations using OECD STAN database and Google Trends data.

Recent research by Ferrara and Simoni (2022[10]) shows that Google Trends data is mainly useful for gauging aggregate GDP growth in the absence of other official data. In the context of aggregate GDP, the data lag typically ranges from one to three months across most countries. However, at the sectoral level, the data lag often extends well beyond twelve months.

To empirically examine whether search indexes lose their explanatory power when other official data becomes available, a two-stage model was developed. The model incorporates more readily available aggregate GDP growth rates. Specifically, the first stage of the model uses these growth rates as the dependent variable, and its residuals serve as the dependent variable for the second stage, which employs an artificial neural network.

The results of this approach are reported in the first column of Table 5.2. The findings indicate that the model's performance is notably inferior to that of a neural network directly trained on the ICT sector's growth rates. This underscores the continued relevance of Google Trends data for assessing the ICT sector growth, even when other official datasets become accessible.

Other machine learning methodologies were also assessed, with the gradient boosting model emerging as the worst performer. This model operates on two fundamental principles: boosting and gradient descent. Boosting is an iterative technique that aggregates weak learners to form a strong learner. On the other hand, gradient descent is an optimisation algorithm that seeks to minimise a specified function through iterative steps in the direction of steepest descent. In this context, the gradient boosting model serves as a specialised implementation that merges the principles of both boosting and gradient descent.

Surprisingly, the third-best performing model is the autoregressive model (AR1), outperforming the machine learning gradient boosting model. The autoregressive model is a simple statistical model that leverages the relationship between an observation and several lagged observations. In this case, the

autoregressive model is a first-order autoregressive model, which means that it only considers the immediately preceding observation. Given its simplicity, the autoregressive model's performance is remarkably strong.

However, the model's performance is still inferior to that of the neural network, which is the best performing model. This is likely because the neural network can capture more complex relationships between the ICT sector's growth rates and Google Trends data. For this reason, the nowcasting model uses a neural network.

6 Choosing hyperparameters

After selecting a machine learning algorithm, the next critical decision involves choosing the network's hyperparameters. Unlike model parameters, which are learned during training, hyperparameters are pre-set values that govern the learning process[1]. To eliminate simply relying on intuition, an automated grid search was employed to systematically explore the hyperparameter space. The grid search uses the RMSE as the evaluation metric to identify optimal parameters. The number of layers, the learning rate, and the activation function were all chosen through an automatised hyperparameter search.

Hidden layers

The first hyperparameter to be chosen is the number of hidden layers which is the number of layers between the input and output layers. In the present model, the input layer corresponds to two different elements. The first element is Google search indexes corrected following the statistical methods described above. The second element is composed by country fixed effects. Country fixed effects are simply dummy variables that account for any country specific effect that is not captured by Google Trends data. The output layer is the growth rate of the ICT sector in the current year.

The number of hidden layers is a key hyperparameter because it determines the complexity of the model. A model with too few hidden layers will not be able to capture the complexity of the data. A model with too many hidden layers will overfit, making it less generalisable to new data.

To choose their optimal number, the model was trained using different numbers of hidden layers. Based on their performance on the validation data, models with one and two hidden layers performed the best. The two next sections discuss the performance of models with one and two layers respectively.

One hidden layer

For any number of hidden layers, a second modelling decision must be made. This decision regards the number of neurons in each hidden layer. Neurons are the units that make up a neural network. Each neuron is connected to the neurons in the previous layer. As indicated above, a model with too few neurons will not be able to capture the complexity of the data and a model with too many neurons will overfit.

The performance of models with one hidden layer according to the number of neurons is included in Table 6.1. The table shows that the best performance is obtained with 750 neurons. For that number of neurons, the model has a RMSE of 2.58 for the validation data. The table also shows that the performance of the model deteriorates when the number of neurons is increased beyond 800. This suggests that the model is overfitting when the number of neurons is increased beyond 800.

Table 6.1. RMSE for models with one hidden layer

by the number of neurons

Hidden Layers	RMSE Training Data	RMSE Validation Data
50	3.31	2.99
100	3.57	2.94
150	3.09	3.43
200	2.68	3.67
250	2.75	3.15
300	2.91	3.12
350	3.42	2.95
400	3.38	2.97
450	2.74	3.69
500	2.84	3.44
550	3.06	3.43
600	3.56	2.76
650	3.34	3.22
700	3.16	3.19
750	3.52	2.58
800	2.92	2.75
850	3.11	2.95
900	2.89	3.62
950	3.33	3.01
1 000	3.22	2.93
1 050	3.06	2.93
1 100	3.23	2.87

Note: The table shows the RMSE for the training and validation data for single hidden layers with different numbers of neurons.
Source: Author's calculations using OECD STAN database and Google Trends data.

Two hidden layers

The performance of models with two hidden layers according to the number of neurons in each layer is included in Table 6.2. The table shows that the best performance is obtained with 2 400 neurons in the first hidden layer and 800 neurons in the second hidden layer. For that number of neurons, the model has a RMSE of 2.98 for the validation data and 2.73 for the training data.

Another model that performs well is the model with 120 neurons in the first hidden layer and 40 neurons in the second hidden layer. For that number of neurons, the model has a RMSE of 2.85 for the validation data and 2.82 for the training data.

After assessing the performance of these two models by using various alternative training datasets, the model with 2 400 neurons in the first hidden layer and 800 neurons in the second hidden layer was chosen. On average, this model had a lower RMSE for the alternative validation datasets than the model with 120 neurons in the first hidden layer and 40 neurons in the second hidden layer. As a result, the nowcasting model uses a neural network with 2 400 neurons in the first hidden layer and 800 neurons in the second hidden layer.

Table 6.2. RMSE for models with two hidden layers

by the number of neurons

Hidden Layers	RMSE Training Data	RMSE Validation Data
100, 50	2.39	3.18
100, 33	2.58	3.46
120, 40	2.85	2.82
150, 50	2.89	3.32
200, 67	3.25	2.86
300, 100	2.59	3.30
600, 200	2.76	3.09
800, 200	2.59	3.45
1 000, 400	1.97	3.36
1 000, 600	2.81	3.26
1 200, 400	3.27	2.87
1 800, 600	2.19	3.77
1 800, 800	1.85	3.80
2 400, 800	2.98	2.73

Note: The table shows the RMSE for the training and validation data for double hidden layers with different numbers of neurons.
Source: Author's calculations using OECD STAN database and Google Trends data.

Activation function

A last modelling choice concerns the activation function. The activation function is a mathematical function that determines the output of a neural network. The activation function is applied to the weighted sum of the inputs of a neuron.

To determine the best option, the model was trained using the four most used activation functions: rectified linear unit (ReLU), identity, logistic and tanh. The identity activation function is the simplest activation function as it is simply a linear function. It returns the value of the input when the neurons are activated. On the contrary, ReLU is a more complex function. It is, however, the most used activation function in deep learning models. The ReLU function is a piecewise linear function that outputs the input directly if it is positive, otherwise it outputs zero. In many different cases, it has proven to offer the best performance. It has also the advantage of being computationally efficient.

Table 6.3 shows the performance of the model using each activation function. The table shows that the best performance is obtained with the ReLU and tanh activation functions. When evaluating the model using alternative validation datasets, the model with the ReLU activation function performed better than the model with the tanh activation function. As a result, the nowcasting model uses the ReLU activation function.

Table 6.3. RMSE for different activation functions

Activation Function	RMSE Training Data	RMSE Validation Data
ReLU	2.98	2.73
identity	2.84	4.19
logistic	3.45	2.89
tanh	2.62	3.07

Note: The table shows the RMSE for the training and validation data for the different activation functions.
Source: Authors' calculations using OECD STAN database and Google Trends data.

7 Standard errors

A last choice concerns the computation of standard errors. Following the agnostic principle guiding the nowcasting model, standard errors are computed using a bootstrapping procedure that does not rely on any specific assumptions about the distribution of the data. More precisely, the bootstrapping procedure follows Effron's (1990[11]) percentile approach. This approach is also used by OECD's weekly tracker (OECD, n.d.[4]).

For this purpose, 2 000 random samples with replacement are drawn from the data. Then, the model is retrained using each new sample. Next, standard errors are ordered to keep 90% confidence intervals. The bootstrapping procedure performs acceptably well. However, it is not perfect. In particular, in a few cases the bootstrapping procedure does not yield centred confidence intervals.

Uncentred bootstrapped standard errors are likely pointing to the fact the distribution of the estimator is not normal. This in turn might signal that the neural network estimator is biased. While having an unbiased estimator is always desirable, there is a bias/variance trade-off that must be considered. In other words, despite the probable bias of the neural network estimator, it still performs well by exhibiting a lower variance.

In this case, there might be a bias between the estimates and the true population parameter. This bias is a plausible explanation for uncentered confidence intervals. Note that given the nonlinear nature of the network, it is generally impossible to characterise the probability distribution of the estimates (Javanmard and Montanari, 2014[12]). This is in line with recent research on the bias of machine learning algorithms (see for instance (Chernozhukov et al., 2018[13]) and (Farrell, Liang and Misra, 2021[14])).

The computation of standard errors is an area where future improvements could be introduced to the nowcasting model.

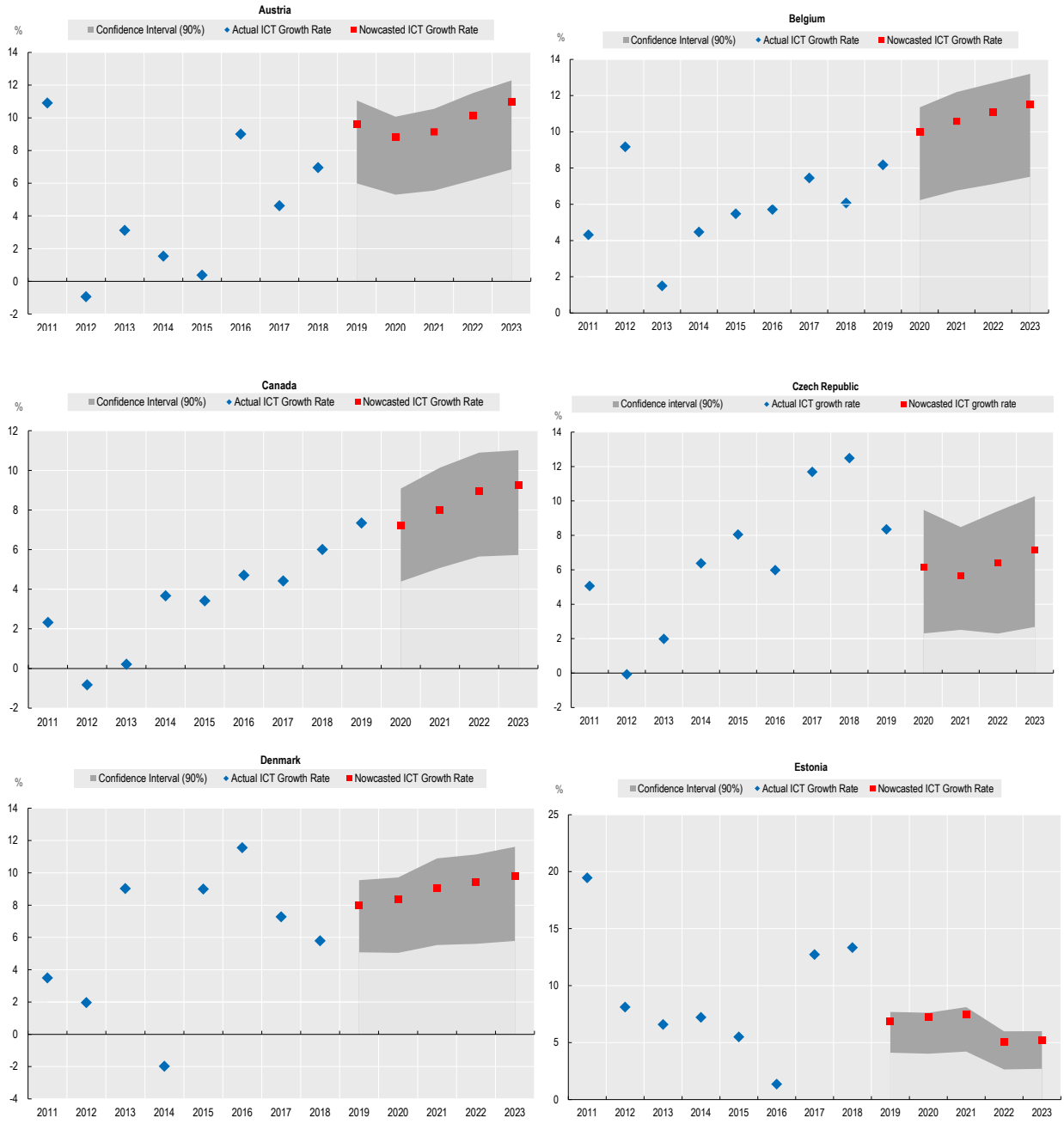
8 Brief overview of the nowcasting results

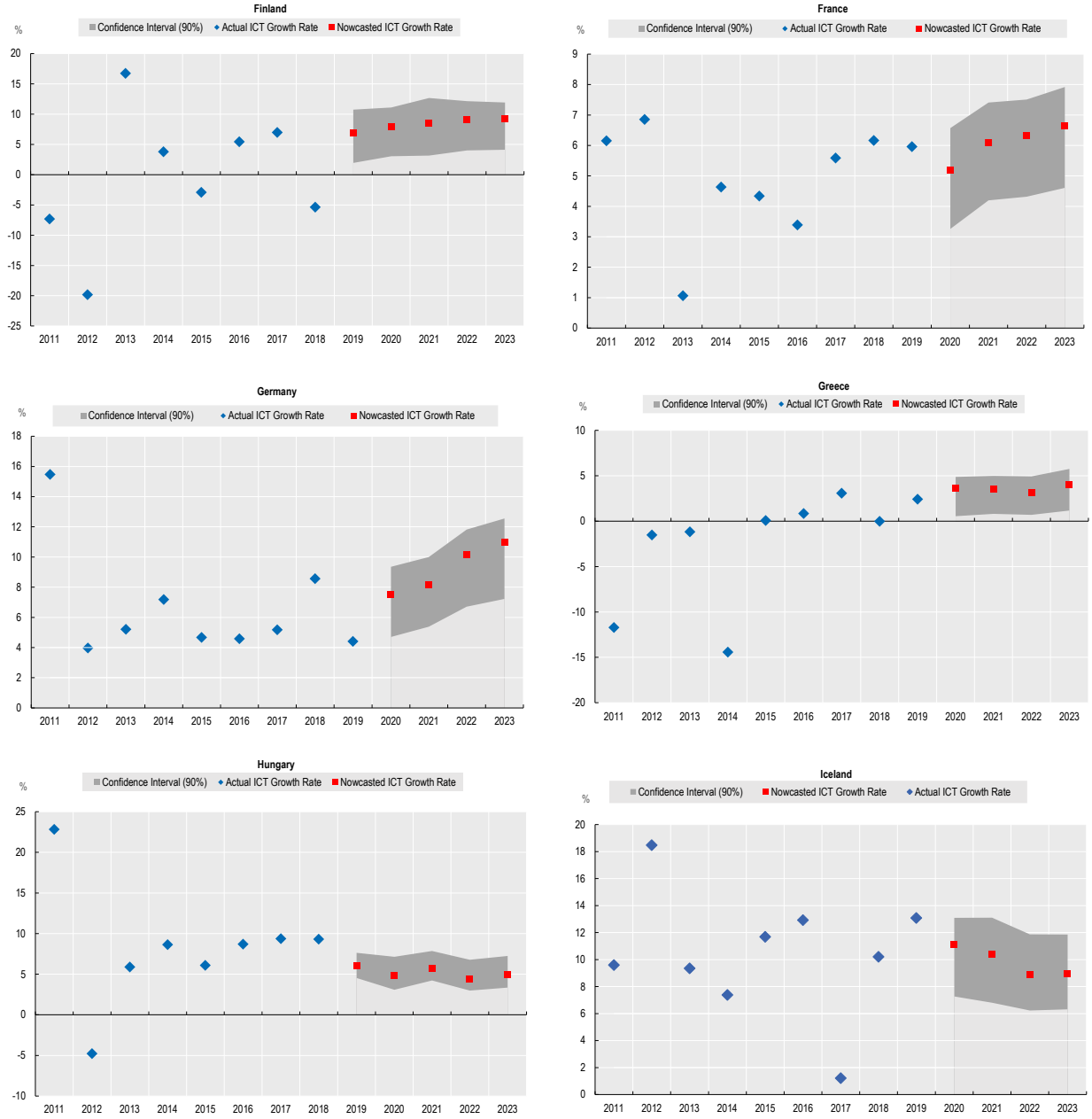
The nowcasting model provides estimates of economic performance for years where observed data is missing. For most countries, this pertains to the years 2020 to 2023. However, for some countries, the OECD's STAN database has data only up until 2018; for these, the model estimates the ICT sector's growth from 2019 to 2023.

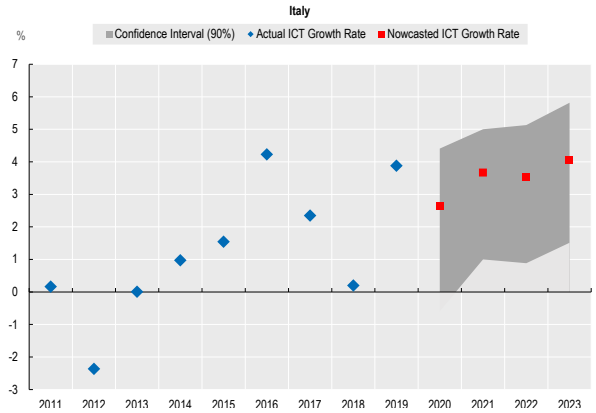
More specifically, the model generates point estimates of the growth rate of the ICT sector. These estimates serve as the basis for analysing this sector's economic performance across OECD countries in Chapter 1 of the O*ECD Digital Economy Outlook 2024 (Volume 1)* (OECD, 2024[1]). In addition to point estimates, the chapter also offers 90% confidence intervals, which should be considered when interpreting the point estimates.

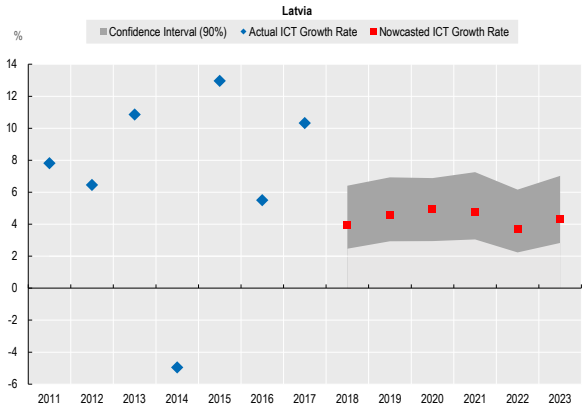
Figure 8.1 presents both observed and nowcasted ICT growth rates for all OECD countries for which all required data are available[2]. The results show that COVID-19 marked an end to an era of sustained increases in growth rates that started shortly after the global financial crisis. However, while growth rates decrease, the growth of the sector nevertheless remains strong: in 2020, the average growth rate of the ICT sector in OECD countries was 6.6%. By 2021, in most OECD countries, the impact of the COVID-19 crisis is not visible anymore with growth rates achieving their historical maxima in 2023. These results must however be nuanced when looking at point estimates within respective confidence intervals[3]. However, even when considering confidence intervals, the nowcasting results show that the COVID-19 crisis only had a moderate impact on the ICT sector.

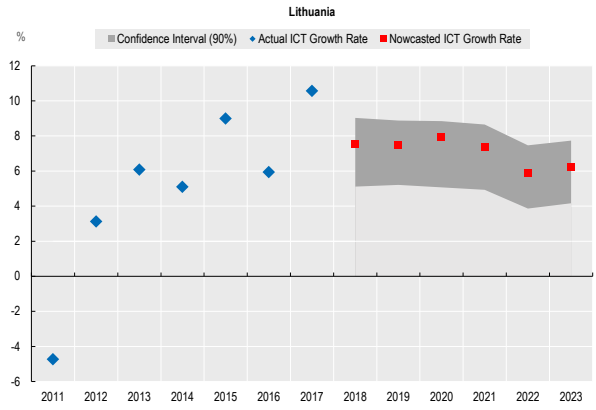
Figure 8.1. Observed and predicted ICT sector growth rates, 2011-23

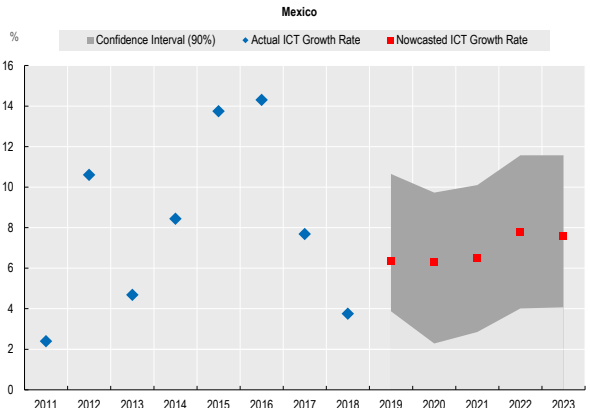


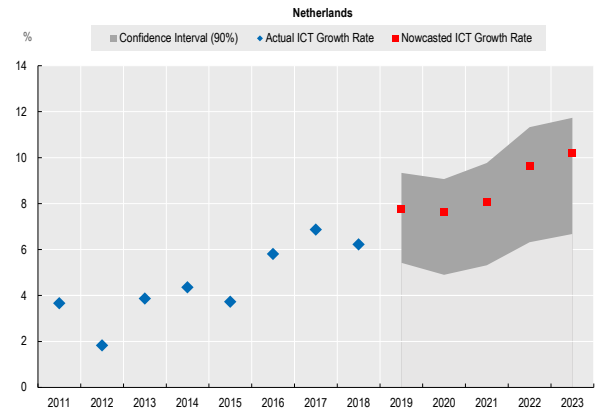

Finland
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
%
France
Germany
Greece
Hungary
Iceland

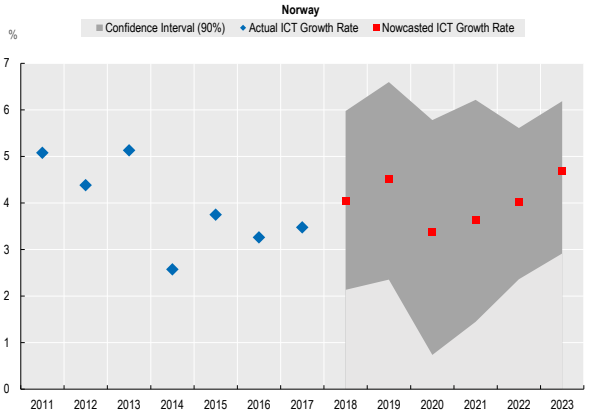

Italy
%
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
7
6
5
4
3
2
1
0
-1
-2
-3
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

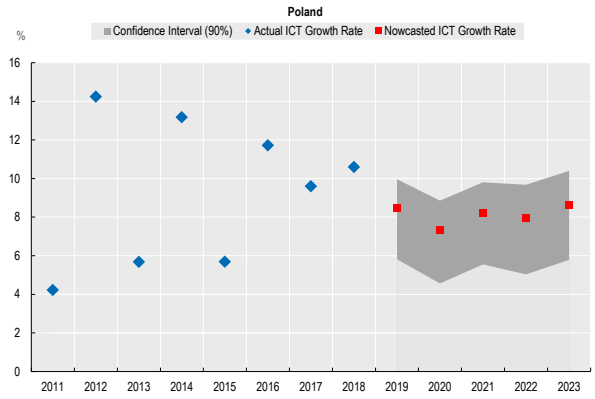

Latvia
%
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
14
12
10
8
6
4
2
0
-2
-4
-6
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

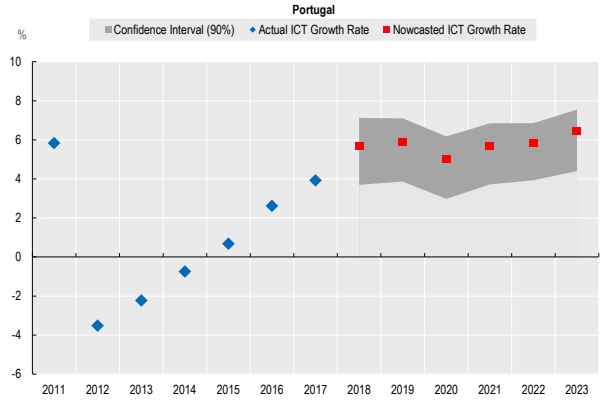

Lithuania
%
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
12
10
8
6
4
2
0
-2
-4
-6
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

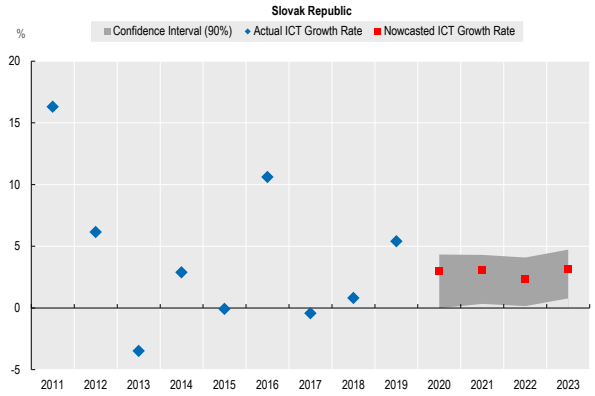

Mexico
%
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
16
14
12
10
8
6
4
2
0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

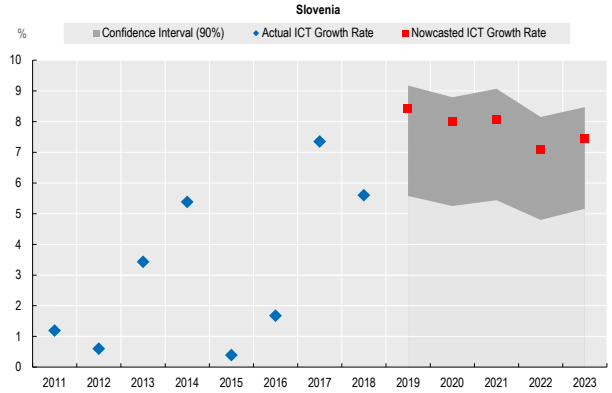

Netherlands
%
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
14
12
10
8
6
4
2
0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

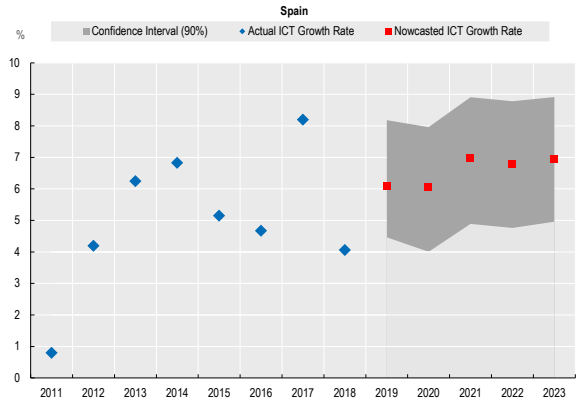

Norway
%
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
7
6
5
4
3
2
1
0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

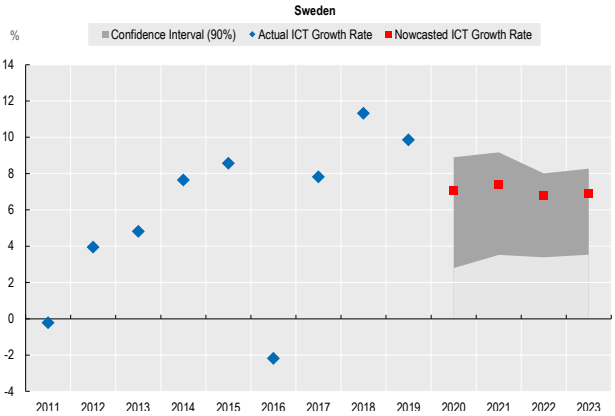

Poland
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
%

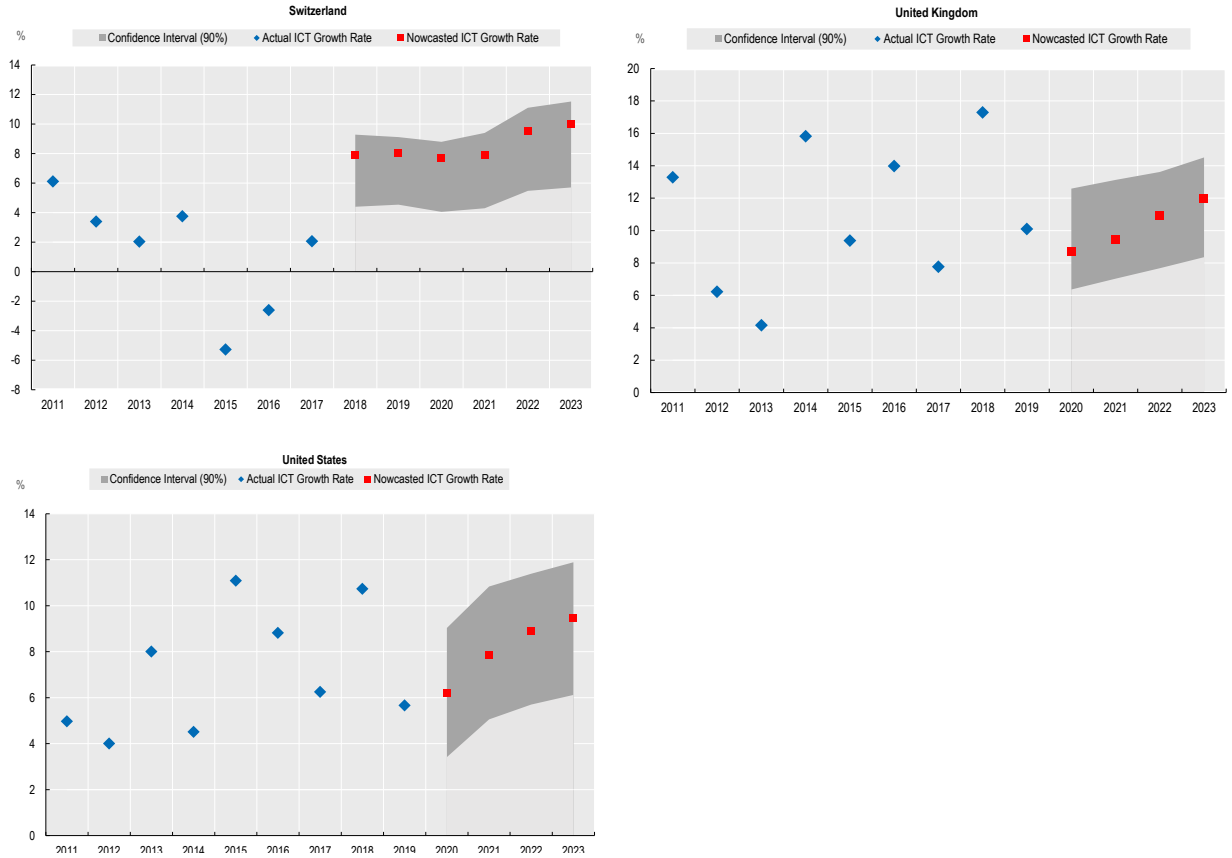

Portugal
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
%


Slovak Republic
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
%


Slovenia
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
%


Spain
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
%


Sweden
Confidence Interval (90%)
Actual ICT Growth Rate
Nowcasted ICT Growth Rate
%



Notes: Observed OECD STAN growth rates ("Actual ICT Growth Rate") are represented by blue diamonds while nowcast estimates are represented by red squares.
Source: Author's calculations using OECD STAN database and Google Trends data.

9 Concluding remarks

In a rapidly evolving digital landscape, accurate and timely economic data is crucial for both evaluating and designing sound economic policies. Traditional data sources often fall behind in their capacity to provide up-to-date insights of sectors as dynamic as the ICT sector. This technical paper details the strategy implemented in Chapter 1 of O*ECD Digital Economy Outlook 2024 (Volume 1)* (OECD, 2024[1]) to fill this gap by leveraging Internet search data as an innovative source for nowcasting ICT sector growth rates.

This technical report provides a comprehensive overview of the methodology used to generate the nowcasting model's results, including the statistical and machine learning techniques employed. In this regard, the paper discusses the challenges associated with using Google Trends data, such as downward trends, sampling noise, and seasonality. It then outlines the strategies used to address these challenges, by the implementation of a Hodrick-Prescott filter and Lowess smoother, repeated sampling, and the choice of the machine learning algorithm. Finally, the paper presents the results of the nowcasting model, which indicate that the ICT sector showed remarkable resilience in the face of the COVID-19 pandemic.

It is worth noting that further methodological improvements are still possible. For instance, the nowcasting model could be enhanced by incorporating additional data sources. Additionally, the model could be improved by testing additional machine learning techniques, such as recurrent neural networks, which are particularly well-suited for time series data. Finally, the computation of standard errors could be improved by developing new statistical procedures.

References

Bantis, C. et al. (2023), "Forecasting GDP growth rates in the United States and Brazil using Google Trends", *International Journal of Forecasting*, Vol. 39/4, pp. 1909-1924, https://doi.org/10.1016/j.ijforecast.2022.10.003. [5]

Chernozhukov, V. et al. (2018), "Double/debiased machine learning for treatment and structural parameters", *The Econometrics Journal*, Vol. 1/1, pp. C1–C68, https://doi.org/10.1111/ectj.12097. [13]

Choi, H. and H. Varian (2012), "Predicting the present with Google Trends", *Economic record*, Vol. 88, pp. 2-9, https://doi.org/10.1111/j.1475-4932.2012.00809.x. [2]

Efron, B. (1990), "More efficient bootstrap computations", *Journal of the American Statistical Association*, Vol. 85/409, pp. 79-89, https://doi.org/10.1080/01621459.1990.10475309. [11]

Eichenauer, V., R. Indergand and I. Martínez (2022), "Obtaining consistent time series from Google Trends", *Economic Inquiry*, Vol. 60/2, pp. 694-705, https://doi.org/10.1111/ecin.13049. [9]

Farrell, H., T. Liang and S. Misra (2021), "Deep neural networks for estimation and inference", *Econometrica*, Vol. 89/1, pp. 181-213, https://doi.org/10.3982/ECTA16901. [14]

Ferrara, L. (2022), "When are Google data useful to nowcast GDP? An approach via preselection and shrinkage", *Journal of Business & Economic Statistics*, Vol. 41/4, pp. 1188-1202, https://doi.org/10.1080/07350015.2022.2116025. [10]

Götz, T. and T. Knetsch (2019), "Google data in bridge equation models for German GDP", *International Journal of Forecasting*, Vol. 35/1, pp. 45–66, https://doi.org/10.1016/j.ijforecast.2018.08.001. [8]

Heikkinen and Joni (2019), *Nowcasting gdp growth using google trends*, https://jyx.jyu.fi/handle/123456789/66363. [7]

Horvát, P. and C. Webb (2020), "The OECD STAN Database for industrial analysis: Sources and methods", *OECD Science, Technology and Industry Working Papers*, No. 2020/10, OECD Publishing, Paris, https://doi.org/10.1787/ece98fd3-en. [3]

Javanmard, A. and A. Montanari (2014), "Confidence intervals and hypothesis testing for high-dimensional regression", *The Journal of Machine Learning Research*, Vol. 15/1, pp. 2869-2909, https://www.jmlr.org/papers/volume15/javanmard14a/javanmard14a.pdf. [12]

Kohns, David and A. Bhattacharjee (2023), "Nowcasting growth using google trends data: A bayesian structural time series model", *International Journal of Forecasting*, Vol. 39/3, pp. 1384–1412, https://doi.org/10.1016/j.ijforecast.2022.05.002. [6]

OECD (2024), *OECD Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier*, OECD Publishing, Paris, https://doi.org/10.1787/a1689dc5-en. [1]

OECD (n.d.), *OECD Economics Department Working Papers*, OECD Publishing, Paris, https://doi.org/10.1787/18151973. [4]

Endnotes

[1] Please see: https://cloud.google.com/blog/products/ai-machine-learning/hyperparameter-tuning-cloud-machine-learning-engine-using-bayesian-optimization.

[2] STAN does not include valued added for the ICT sector for: Australia, Chile, Colombia, Costa Rica, Ireland, Israel, Japan, Korea, Luxembourg, New Zealand and Türkiye.

[3] Sectoral growth rates display a considerably larger variance than total GDP growth rates. This feature of sectoral data explains to a large extent the magnitude of the nowcasting model confidence intervals (90%) compared to those using a similar methodology for total GDP growth rates (95%) (OECD, n.d.[4]).