INTERNATIONAL MONETARY FUND

Overcoming Data Sparsity: A Machine Learning Approach to Track the Real-Time Impact of COVID-19 in Sub-Saharan Africa

By Karim Barhoumi, Seung Mo Choi, Tara Iyer, Jiakun Li, Franck Ouattara, Andrew Tiffin, and Jiaxiong Yao

2022
MAY

WORKING PAPER

WP/22/88

IMF Working Paper

African Department

Overcoming Data Sparsity: A Machine Learning Approach to Track the Real-Time Impact of COVID-19 in Sub-Saharan Africa

Prepared by Karim Barhoumi, Seung Mo Choi, Tara Iyer, Jiakun Li, Franck Ouattara, Andrew Tiffin, and Jiaxiong Yao

Authorized for distribution by Papa N'Diaye
May 2022

ABSTRACT: The COVID-19 crisis has had a tremendous economic impact for all countries. Yet, assessing the full impact of the crisis has been frequently hampered by the delayed publication of official GDP statistics in several emerging market and developing economies. This paper outlines a machine-learning framework that helps track economic activity in real time for these economies. As illustrative examples, the framework is applied to selected sub-Saharan African economies. The framework is able to provide timely information on economic activity more swiftly than official statistics.

Overcoming Data Sparsity: A Machine Learning Approach to Track the Real-Time Impact of COVID-19 in Sub-Saharan Africa

Prepared by Karim Barhoumi, Seung Mo Choi, Tara Iyer, Jiakun Li, Franck Ouattara, Andrew Tiffin, and Jiaxiong Yao[1]

TABLE OF CONTENTS

I. INTRODUCTION

A lack of adequate and timely data often prevents policymakers from monitoring economic activity in real time. This has been a major challenge for policymakers during the COVID-19 pandemic as government action is needed urgently. Many emerging market and developing economies have limited institutional capacity and do not necessarily compile and publish timely macroeconomic statistics. Official national accounts data on GDP, for example, are released with substantial delays. In such circumstances, the lack of information could undermine the quality of policy operations, especially during a crisis.

This study seeks to bridge the gap between policymaking and data availability by developing a framework to track real-time economic activity in emerging markets and developing economies. The paper focuses on countries in sub-Saharan Africa, where data scarcity issues are pervasive. Economic activity is tracked using a "nowcasting" framework that extracts signals from indicators that are available earlier than the official GDP series. These signals allow for the real-time prediction of economic activity, often months or quarters before official GDP statistics are released. The framework draws on a broad range of increasingly popular machine-learning techniques. Machine learning has gained attention as an expanding sub-field of applied computational statistics and is ideally suited to the prediction challenge that lies at the core of the nowcasting problem. This is the first study to apply this methodology to a broad range of sub-Saharan African countries.

As a flexible framework that focuses primarily on accurate predictions, machine-learning algorithms have often proven more effective than traditional econometric methods. Machine learning models frequently employ techniques that are familiar to most economists; indeed, most machine-learning textbooks start with simple ordinary least squares (OLS) regression (see Appendix). However, the focus of machine-learning models is somewhat different from traditional econometric models. Instead of exploring issues of identification and causality, the primary focus of machine learning is on producing more precise out-of-sample *predictions*. To this end, a machine-learning framework will exploit historical, often non-linear, statistical patterns in the data, without necessarily assuming that these relationships are known in advance (see Appendix). Applied to the nowcasting problem, a machine-learning framework will seek to extract reliable signals from a large set of noisy high-frequency indicators, capturing the co-movement between these indicators and GDP.

The paper is structured as follows. Section II discusses the literature and places this study in context. Section III provides an overview of the framework used to nowcast GDP growth in sub-Saharan Africa. Section IV discusses the framework's predictions during the COVID-19 pandemic in sub-Saharan Africa. Section V concludes. The appendix provides an overview of machine-learning methods.

II. RELATED LITERATURE

The concept of nowcasting has long been a topic of interest among policymakers and has gathered increasing attention, particularly over the past 10–15 years. Following Giannone and others (2008), nowcasting models have been adopted at several central banks in advanced economies (Richardson and others 2021). Recent studies also develop nowcasting models for emerging markets and developing economies, including Turkey (Solmaz and Sanjani 2015), Lebanon (Tiffin 2016), India (Iyer and Gupta 2019), and others (Marini 2016; Narita and Yin 2018). This paper contributes to the literature by expanding the application to the data-sparse environment of sub-Saharan Africa.[2]

Machine-learning algorithms have often performed relatively well when quickly capturing sharp turning points in GDP growth. Jung and others (2018) test for the robustness of machine-learning forecasts in historical crises and find that, although the accuracy of their machine-learning framework deteriorated during crises, it nonetheless remained superior to other econometric approaches. Going further, Hu and others (2019) distinguish between crisis periods and non-crisis periods to find that machine-learning algorithms outperform linear regressions in both samples.

Machine learning provides an alternative to the dynamic factor model methodology commonly used in the nowcasting literature. Factor-based models extract a small set of latent factors from a large set of indicators by exploiting the co-movement among variables (for example, Giannone and others 2008; Barhoumi and others 2012; Bok and others 2017; and Iyer and Gupta 2019). This approach has often performed well. However, even though the factors are generally able to efficiently summarize the information available in the indicator set, the models are not always focused on those variables that might individually be better predictors of the output indicator.

The machine learning approach has been often studied with non-traditional data sources. For instance, Cerdeiro and others (2020) use real-time maritime data to nowcast world seaborne trade volume to show how international trade has been affected by the COVID-19 pandemic. Similarly, Carton and others (2020) forecast short-term international trade using the Society for Worldwide Interbank Financial Telecommunication (SWIFT) messages and other high-frequency data. Machine learning has also been used to predict credit-default swap (CDS) spreads (Hu and others 2019).

---

[2] Buell and others (2021) discuss potential projection tools without applying them to the COVID-19 period.

III. NOWCASTING FRAMEWORK

This new framework has three stages: 1) selecting predictors, 2) selecting the best model out of competing models (the "horserace"), and 3) nowcasting the GDP using the best model.

A. Stage 1: Selecting Predictors

Predictors are selected based on three criteria. First, predictors should be historically related (linearly or non-linearly) to GDP growth.[3] Second, predictors should be released in a timely fashion, ideally well before the release of GDP figures. Third, data for predictors should be available for a sufficiently long time period, ideally matching the length of the available GDP series.[4] Figure 1 provides an example of two key indicators that could be used to nowcast Nigeria's GDP growth. The first is the global oil price, which moves in sync with GDP growth in Nigeria. Data on the oil price becomes available quickly (almost in real time), and given Nigeria's status as an oil exporter, could be a useful predictor. However, a second indicator, comprising of the year-over-year change rate in Nigeria's stock price index, is less promising as a predictor because there is no clear correlation with Nigeria's GDP growth.

Figure 1: Selecting Predictors: Example

The global oil price could be a good predictor as it moves together with Nigeria's GDP growth, the data become available quickly, and the data go back to an earlier period. However, the year-on-year change rate in Nigeria's stock price index is less promising as the relationship with GDP growth is not clear.
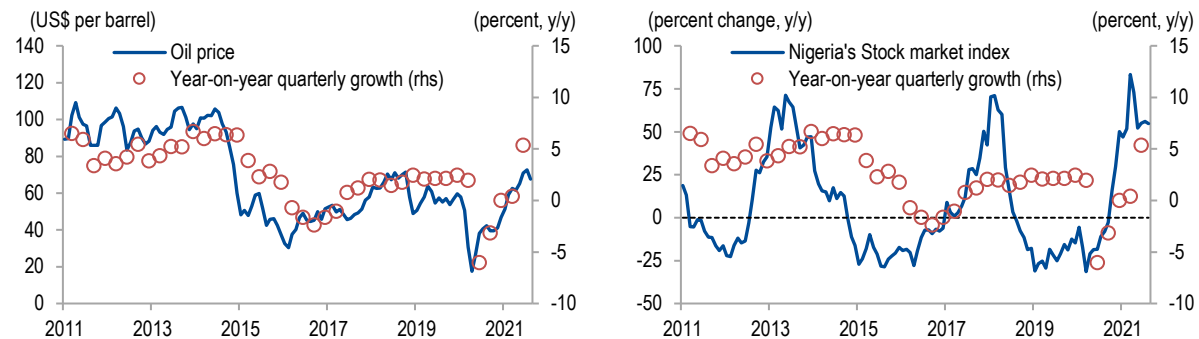
Sources: Haver; and IMF staff calculations.
Note: rhs = right-hand scale; y/y = year-on-year.

The most relevant predictors depend on expert knowledge, careful statistical analysis of trends, and the country context. Tourist arrivals would be more appropriate for countries such as Cabo Verde, Mauritius, and Seychelles. The country's economic survey indicators (such as the

---

[3] Useful predictors may not always have high correlation with GDP growth since the relationship may not be linear. Machine learning has the advantage of capturing both linear and non-linear relationships.

[4] For example, one such indicator is the Google mobility indicator, which may be related to economic activity and the impact of lockdown measures. However, the time series for this is relatively short, starting from early 2020, leading to missing data issues. Incorporating such indicators is a useful topic for future research, including developing more efficient algorithms to deal with missing data and using non-macroeconomic indicators for nowcasting.

purchasing managers index [PMIs] and consumer confidence index), vehicle sales, imports, exports, real effective exchange rate (REER), fiscal revenue, and the global prices of commodities that the country exports heavily, if available, can be considered as potential predictors. Nontraditional indicators, such as nighttime lights, nitrogen dioxide (NO_2) emissions, and Google search volume indices, also can be considered. Finally, a COVID-19 dummy equal to one during the COVID-19 crisis, or estimating the model with a structural break during the crisis, could help capture the potential impact of the pandemic, which would not be captured by other predictors.

Useful predictors are not limited to those compiled within the country. Key economic indicators from neighboring economies or major trading partners may also be considered as predictors. This is because they could reflect cross-country spillovers or common regional or global factors (such as financial conditions) affecting different countries simultaneously. For example, PMIs in China, Nigeria, South Africa, and other major regional or global economies, could be useful. The stock market index in South Africa could provide a useful market signal for other sub-Saharan African countries, even if there are not strong financial linkages between these countries, because South Africa's financial markets could reflect financial conditions and risk appetite in the region. Furthermore, the Brent crude oil price can be considered. For oil exporters and for countries that are closely linked to them, the oil price often co-moves with GDP. For oil importers, the opposite may be true. The industrial materials index compiled by the Foundation for International Business and Economic Research (FIBER) also can be considered since it reflects global industrial demand (although it also co-moves with the oil price so may be less useful if the oil price is already included as a predictor).

For the particular purpose of nowcasting GDP growth in sub-Saharan African countries, including too many predictors does not always help. The time horizon for GDP data in these countries tends to be limited. For example, if the country has about 10 years of quarterly GDP data (as is the case for several countries in the region), only about 40 data points exist for model estimation. In this case, adding a predictor may lead to the deterioration of model performance if it is not clearly correlated with GDP growth. For example, the euro area's PMI could be viewed as a potentially useful predictor for a sub-Saharan African country's GDP as there is likely a positive relationship between economic activities in the euro area and sub-Saharan Africa (for example, reflecting trade in goods and services). However, during our sample period of the recent 10 years or so, the euro area and sub-Saharan Africa faced large negative shocks during different periods. That is, the euro area had a financial crisis around 2012 (during which the PMI was low) while sub-Saharan Africa faced a large impact from the global commodity price shock during 2015–16. As a result, for some sub-Saharan African countries, the euro area's PMI can be seen by our framework as moving in the *opposite* direction of the country's GDP growth during the sample period. Therefore, if the euro area's PMI is included, the framework could take the euro area's economic recovery from the COVID-19 crisis as a negative development for some sub-Saharan African countries.

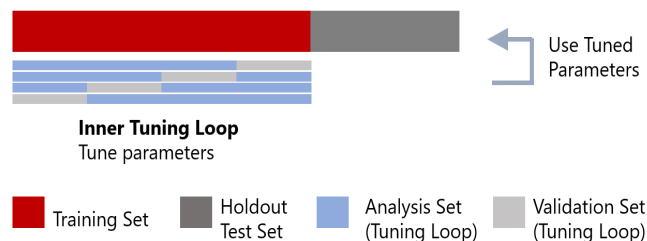## B. Stage 2: Selecting the Best Model ("Horseracing")

Having selected the predictors for each country, the next stage in the framework is to evaluate a range of potential machine-learning models based on their out-of-sample predictive performance. For each machine learning model applied to the data, the first 85–95 percent of data is used as a training dataset, and the remaining 5–15 percent is set aside as a holdout dataset to evaluate the out-of-sample performance of the model. A separate horserace is run for each country, so that the "best" model will often differ from country to country.

Before being evaluated on the holdout test set, each model is "tuned" to optimize its likely out-of-sample performance. While details will differ, most models include several "hyperparameters" that shape the model's complexity and performance. These must be set ex-ante and are typically chosen to optimize the model's likely predictive performance. Within each training set, therefore, the framework uses three-fold cross validation to tune the hyperparameters. Cross validation is a sampling technique that essentially conducts several simulated (in-sample) experiments to gauge the likely

Figure 2: Cross Validation

Cross validation conducts several simulated experiments to gauge the likely out-of-sample performance of a particular model. Then, the framework evaluates the performance of more than 30 different types of models on the holdout test set.

Source: IMF staff.

out-of-sample performance of a particular model (Figure 2). With three-fold cross validation, the training set is split into three random folds. For a given set of hyperparameters, the model is trained using two of those folds and the remaining fold is used to measure the model's prediction error (for that set of hyperparameters). With three folds, this experiment can be run using three different combinations, providing three measurements of the model's prediction error. For this study, the framework uses the Root Mean Squared Error (RMSE) as the measurement.[5] The average of these three errors then provides a measure of how well that model is likely to do out of sample. This set of experiments can be conducted for a range of different hyperparameters, which then allows the researcher to choose the set of hyperparameters most likely to lead to the best performance. Once a "tuned" model is determined from the training set, that model is then used to predict GDP growth in the holdout test set.

---

[5] The RMSE, which measures the distance between the actual time series and the predicted values, is commonly used to evaluate how close the predictions are to the data. An alternate metric is the mean absolute error (MAE), which puts relatively less penalty weight on predictions with large errors and, therefore, makes the model less sensitive to them. There are discontinuities in its derivative, which hinder its widespread use. Thus, RMSE remains analytically convenient and the most popular in the literature. RMSE is used both in the training sample to tune hyperparameters and in the hold-out set to choose the algorithm in the horserace.

The nowcasting framework evaluates the performance of more than 30 different types of models on the holdout test set—having already "tuned" each model on the training set.[6] Additionally, there is an ensemble option, which is a second-layer algorithm that combines the predictions of the previous models. While a detailed treatment of these models can be found elsewhere (see, for example, Flach 2012), the Appendix to this paper provides an overview of some of the primary algorithms used in the nowcasting framework. Having evaluated more than 30 different types of models, the winner of the horserace is the one with the best likely out-of-sample performance. The model with the lowest RMSE during the hold-out evaluation period is typically selected.

C. Stage 3: Nowcasting

The framework now deploys the winning model. The best model selected by the horserace is then re-estimated using the entire sample, rather than just the test set. The predictions from this re-estimated model then serve as the basis for the nowcast.

Although these types of machine-learning models do not aim to identify causal relationships, an "interpretable machine-learning approach" can nonetheless be used as a guide for why the chosen model arrives at a particular projection. Some of the machine-learning algorithms, such as gradient boosting, are often considered a "black box" due to the difficulty in understanding how the results were derived. The authors of this paper draw on recent advances in the growing field of "interpretable machine learning" to help unpack the framework's nowcast projections. These include Shapley decompositions.

Shapley values draw on cooperative game theory to fairly estimate the contribution of each predictor to an individual projection. For the chosen model, Shapley values indicate which variables prompted the model to differ from the sample average, even accounting to non-linearities, and will provide a quantitative guide for each variable's relative contribution to the point prediction.[7]

D. Illustrative Examples

As an illustrative example of the predictive ability of the framework, Nigeria is first considered.[8] The predictors include Nigeria's real effective exchange rate; Nigeria's imports from the US and India;

---

[6] The techniques considered include OLS, step model, elastic net, principal component regression, partial least squares regression, multivariate adaptive regression spline, random forest, stochastic gradient boosting trees, support vector machine (linear, polynomial, and radial basis function), relevance vector machine (linear, polynomial, and radial basis function), and gaussian process (linear, polynomial, and radial basis function) and their "variable selection" variants.

[7] Alternatively, local interpretable model-agnostic explanations (LIME) are a local surrogate method that takes a particular prediction, perturbs the data underlying that prediction, employs a simple linear algorithm to fit the perturbed data, and essentially builds a simpler, more interpretable model for the space around the prediction. This also allows for an exploration of each variable's contribution to the chosen prediction.

[8] This country was chosen because its GDP publication is relatively timely. This enables to compare the framework's projections with realized values during the COVID-19 crisis. In addition, the PMI data are available, which are a useful predictor.

PMIs in Nigeria, South Africa, and China; stock price indices in South Africa (as an indicator reflecting regional financial conditions); the oil price; the FIBER industrial materials index; and the COVID-19 dummy. The models are estimated from the first quarter of 2011, while 90 percent of the sample is used for training purposes.[9]

The horseracing results (in Stage 2) indicate that the machine-learning algorithms generally perform better than competing parametric models. The out-of-sample RMSE is the lowest for "Stochastic Gradient Boosting Trees" among the models considered, including the OLS. The runner-up models are "OLS with Variable Selection" and "Random Forest." Figure 3 shows that projections based on "Stochastic Gradient Boosting Trees" have a lower value of out-of-sample RMSE than some of competing parametric models, such as the OLS, random walk, and an autoregressive model of order 1 (AR(1)).

Figure 3: Nigeria (Out-of-Sample Period): Year-on-Year Rolling Quarterly Real GDP Growth, Data and Projections

The selected model has the out-of-sample RMSE lower than traditional methods such as OLS, random walk, and AR(1).
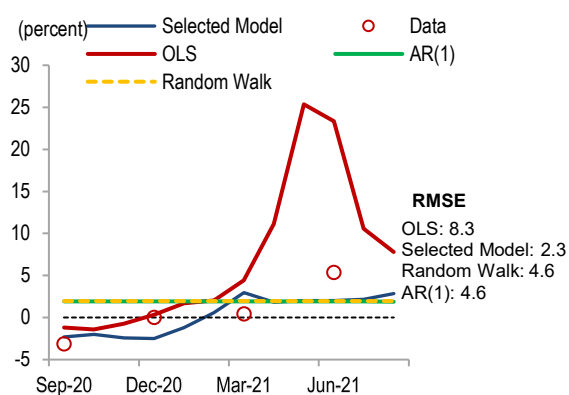
Sources: Nigerian authorities; Haver; and IMF staff calculations.
Note: Random walk and AR(1) projections based on GDP data during the in-sample period only.

Figure 4: Nigeria (All Period): Year-on-Year Rolling Quarterly Real GDP Growth, Data and Projections

The nowcasting framework's projections move closely with data.
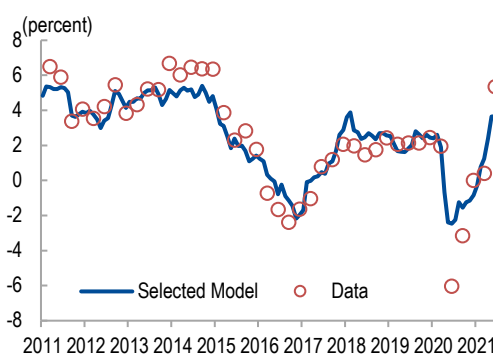
Sources: Nigerian authorities; Haver; and IMF staff calculations.

The final machine learning algorithm chosen based on Step 2 is found to perform well. Figure 4 illustrates that the framework's projections move closely with data. In the second quarter of 2020 (the initial quarter which was fully impacted by the pandemic), the projected contraction (about –2 percent) is narrower than observed (about –6 percent). This suggests that the impact of COVID-19 was more severe than implied by observed predictors. Still, the projection shows an exceptional contraction. Such

---

[9] While there is no clear rule on how long the "training and tuning" period should be (for example, 90 percent), it should include at least one period during the COVID-19 crisis so that the COVID-19 dummy, if included as a predictor, varies over time.

information on projected "turning points" would be useful for policymakers. The selected model projects about 4 percent year-on-year growth ("nowcast") for the quarter ending in August 2021. (Nigeria's third quarter year-on-year growth was later released as 4.0 percent.)

The Shapley decomposition of the nowcast helps to interpret the results (Figure 5). The decomposition for the third quarter of 2021 implies that, compared to Nigeria's average year-on-year growth over the sample (2.5 percent), Kenya's relatively low PMI in the third quarter of 2021 has lowered Nigeria's predicted growth by 0.2 percentage point. Similarly, the COVID-19 dummy lowers the projection by 0.2 percentage point. However, Nigeria's relatively high PMI raises the projection by 0.3 percentage point. South Africa's stock market index, possibly reflecting loose financial conditions, raises the projection by 0.8 percentage point.[10] Adding up these components, Nigeria's nowcast becomes 4.3 percent.

Figure 5: Nigeria: Shapley Decomposition

The Shapley decomposition of projected year-on-year real GDP growth in the quarter of interest is provided.
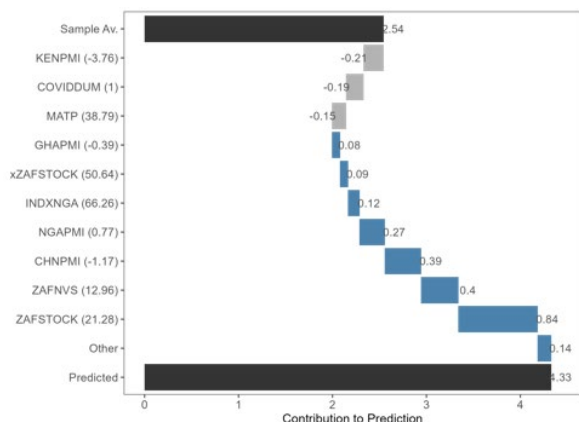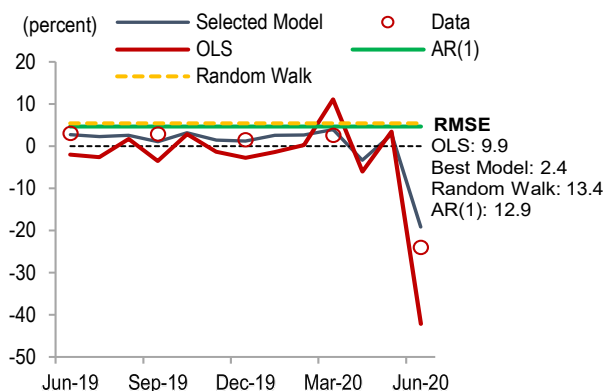

Source: IMF staff calculations.
Note: Sample Av. = Sample average. KENPMI = Kenya's PMI. COVIDDUM = COVID-19 dummy. MATP = Materials price index. GHAPMI = Ghana's PMI. xZAFSTOCK = Stock market index in selected sub-Saharan African countries excluding South Africa. INDXNGA = India's exports to Nigeria. NGAPMI = Nigeria's PMI. CHNPMI = China's PMI. ZAFNVS = South Africa's new vehicles sold. ZAFSTOCK = South Africa's stock market index.

Figure 6: Botswana (Out-of-Sample Period): Year-on-Year Rolling Quarterly Real GDP Growth, Data and Projections

The selected model has the out-of-sample RMSE lower than traditional methods such as OLS, random walk, and AR(1).

Sources: Botswana authorities; Haver; and IMF staff calculations.
Note: Random walk and AR(1) projections based on GDP data during the in-sample period only.

Botswana is considered as another example, projecting GDP growth for the second quarter of 2020 (the first full quarter under the COVID-19 crisis), provided that GDP data for the second quarter of 2020 are not yet released. The predictors include the Google search volume index (SVI) in the "Travel" category for the word "Botswana" (as an indicator for tourism activity in the country), Botswana's real effective exchange rate, stock exchange index, inflation, imports, bank loans to

---

[10] South Africa's stock market potentially reflects regional financial conditions, which also affect Nigeria's economic activity.

households, diamond exports, electricity generation, tax revenue, and GDPs for India and the United States (which are Botswana's major diamond importers). Data start from the second quarter of 2008. "Support Vector Machine" has the lowest out-of-sample RMSE among the models considered. This selected model projects about –19 percent year-on-year growth ("nowcast") for the second quarter of 2020, where the model was run with GDP data only up to the first quarter of 2020 (that is, provided that second quarter 2020 GDP data are not yet released). The actual growth for the second quarter of 2020 was –24 percent. This shows that the "turning point" of GDP growth was relatively well identified through the framework, which could give policymakers more space for early action. The Shapley decomposition suggests that GDP in the United States, GDP in India, and diamond exports explain the nowcast (–19 percent) which is a lot lower than the sample average of GDP growth.

Figure 7: Botswana (All Period): Year-on-Year Rolling Quarterly Real GDP Growth, Data and Projections
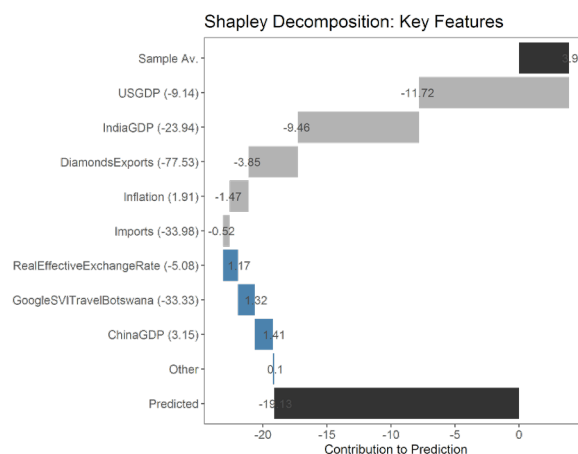
The nowcasting framework's projections move closely with data.
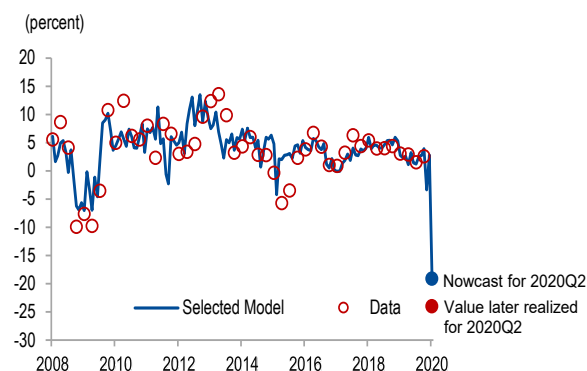
Sources: Botswana authorities; Haver; and IMF staff calculations.

Figure 8: Botswana: Shapley Decomposition

The Shapley decomposition of projected year-on-year real GDP growth in the quarter of interest is provided.

Source: IMF staff calculations.
Note: Sample Av. = Sample average. USGDP = United States' GDP. IndiaGDP = India's GDP. DiamondsExports = Botswana's Diamonds Exports. Inflation = Botswana's Inflation. Imports = Botswana's Imports. RealEffectiveExchangeRate = Botswana's Real Effective Exchange Rate. GoogleSVITravelBotswana = Botswana's Google SVI (Travel). ChinaGDP = China's GDP.
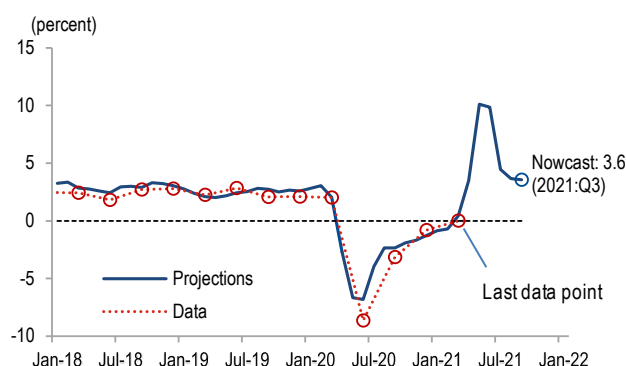
IV. THE COVID-19 CRISIS IN SUB-SAHARAN AFRICA

Sub-Saharan Africa has faced an unprecedented health and economic crisis. The second quarter of 2020 has been particularly damaging with a year-on-year growth rate of −8.5 percent (Figure 9). Since then, sub-Saharan Africa has been experiencing a partial recovery throughout the year. Before the pandemic, sub-Saharan Africa was projected to grow by 3.6 percent in 2020 (October 2020 *Regional Economic Outlook: Sub-Saharan Africa*), which was ultimately brought down to a contraction of −1.7 percent (October 2021 *Regional Economic Outlook: Sub-Saharan Africa*). This contraction, the worst on record, has reversed close to a decade of hard-earned economic and social gains for the continent.

The nowcasting framework tracks real GDP growth. It is applied to project year-on-year real GDP growth in each of the following four economies: (i) South Africa, (ii) Nigeria, (iii) Angola, and (iv) an aggregate of Botswana, Cameroon, Côte d'Ivoire, Ghana, Kenya, Lesotho, Namibia, and Tanzania (weighted by purchasing-power-parity [PPP] GDPs). Then, projections for these four economies are aggregated for sub-Saharan Africa. That is, sub-Saharan Africa is proxied by these 11 countries. These countries have quarterly GDP data available at least from the first quarter of 2010 and account for about ¾ of the region's PPP GDP.[11] The projections closely move together with available GDP growth observations up to the first quarter of 2021.[12]

Figure 9: Sub-Saharan Africa: Year-on-year Rolling Quarterly Real GDP Growth, Data and Projections

After an unprecedented contraction in the second quarter of 2020 at the height of the pandemic, sub-Saharan Africa's economy continued to recover.

Sources: Haver; IMF internal databases; and IMF staff calculations.

Figure 10: Sub-Saharan Africa: Daily COVID-19 Cases

Sub-Saharan Africa has faced three waves of the COVID-19 pandemic, the third one having the largest impact.
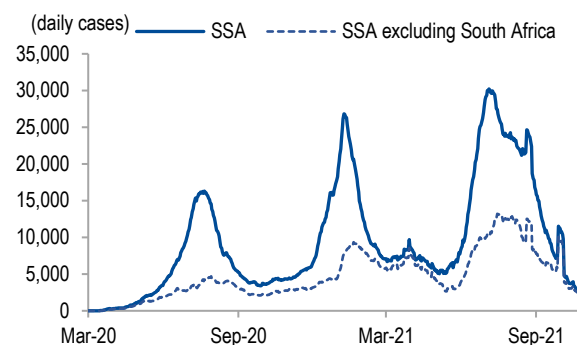
Sources: Center for Systems Science and Engineering, Johns Hopkins University; and IMF staff calculations.
Note: Chart shows seven-day moving averages.
SSA = Sub-Saharan Africa.

---

[11] Ethiopia is excluded due to the unavailability of quarterly GDP data.

[12] Growth in the first quarter of 2021 is assumed to be the same as in the last quarter available for the countries whose data are missing.

**The slowdown in GDP growth in the second quarter of 2020 reflects lockdown measures as well as lower commodity prices and lower external demand.** As sub-Saharan Africa's new COVID-19 cases increased during second quarter of 2020 (Figure 10), voluntary social distancing and lockdown measures, such as quarantine requirements, (Figure 11) contributed to lowered mobility to nonresidential areas (Figure 12). In addition, West Texas Intermediate (WTI) crude oil prices declined from about $60 before the pandemic to below $20 in April 2020. Eurobond spreads also significantly widened at the height of the crisis. The number of flight arrivals dropped substantially during the second quarter of 2020, reflecting worldwide travel restrictions.

Figure 11: Sub-Saharan Africa: Oxford Government Response Stringency Index (0=weakest, 100=strongest)

Containment measures strengthened in March and April 2020, gradually loosened until late 2020, and then remained at a similar level since then.
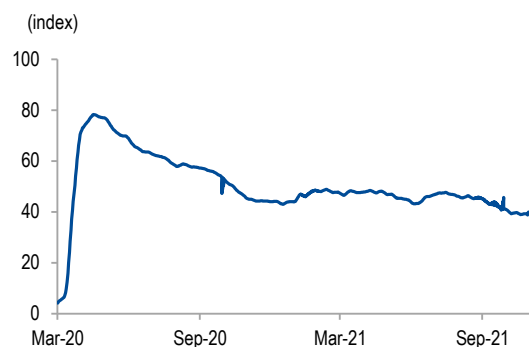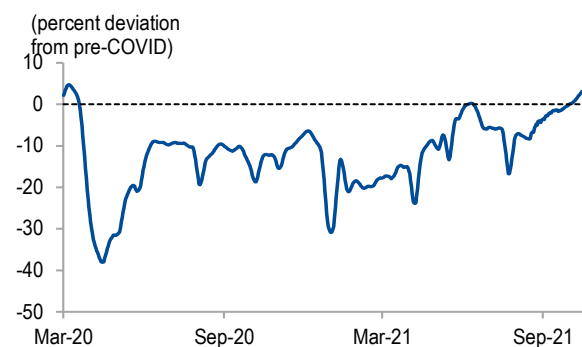

Sources: Oxford COVID-19 Government Response Tracker; and IMF staff calculations.
Note: Figure shows the simple average across countries.

Figure 12: Sub-Saharan Africa: Google Mobility Index for Workplaces

Mobility to workplaces remained lower than pre-COVID-19 until mid-2021. Mobility was lowered temporarily around the second and third waves.

Sources: Google LLC, "Google COVID-19 Community Mobility Reports"; and IMF staff calculations.
Note: Simple average across countries with data available.

As a second quarter 2021 nowcast, the framework predicts high year-on-year growth (about 10 percent), largely reflecting the base effect from the bottom in the second quarter of 2020.[13] As the base effect phases out, growth for the third quarter of 2021 is projected to decline to 3.6 percent, partly reflecting the impact of the third COVID-19 wave in sub-Saharan Africa.

V. CONCLUSION

This study developed an algorithmic framework to track real-time economic activity in sub-Saharan Africa using machine-learning methods. Among the first studies to nowcast GDP growth in sub-Saharan Africa, the findings in this paper indicate that machine-learning algorithms can potentially produce superior nowcasts compared to more traditional regression methods. Using inputs based on

---

[13] As of this writing, Angola, Nigeria, South Africa, and some other countries have released their second quarter 2021 GDP data. However, several countries have not released their data yet.

expert knowledge, analysis of statistical trends, and the specific country context, the nowcasting framework provides a valuable addition to the policymaker's toolkit.

The nowcasting framework for sub-Saharan Africa's GDP growth can enhance regional surveillance. Reflecting monthly releases of high-frequency data, nowcasts would inform IMF staff, as well as policymakers, about the latest status of regional economic development. Furthermore, the nowcast of GDP growth for individual countries (where quarterly GDP data are available) also can be produced on a regular basis as a reference for surveillance. In this case, country-specific information provided by IMF country teams would further improve the nowcasts.

The nowcasting model and results can be used for internal IMF work as well as external engagement with member countries. For example, a nowcasting dashboard could be incorporated in the future based on the framework developed in our study, with regular updates on tracking real-time economic activity. Training sessions at the IMF could be held on the model, and based on resources, presentations of nowcasting results for several individual countries can be made to staff. IMF staff could provide workshops to the authorities of member countries to provide technical expertise and knowledge and to help expand their internal forecasting capacities.

While this study seeks to make a significant contribution in tracking real-time activity in the data-sparse environment in sub-Saharan Africa, there is scope for future research, including improving the coverage of countries that do not publish quarterly GDP data. Countries with sparser data (especially those without quarterly national account statistics) constitute a significant number of sub-Saharan African countries. The lack of quarterly GDP data compiled and published by country authorities could pose statistical challenges in implementing supervised machine-learning algorithms. To tackle this issue, a panel approach with a big data angle can be pursued. Another interesting angle to pursue in the future, especially as some countries lack timely data on macroeconomic series, would be to use non-traditional data, including satellite, mobility, Google trends, flight, and textual data.

VI. REFERENCES

Barhoumi, K., L. Ferrara, and O. Darné. 2012. "Nowcasting the German GDP: A comparison of bridge and factor models." *Journal of Policy Modeling* 34(6): 864–78.

Bok, B., D. Caratelli, D. Giannone, A. Sbordone, and A. Tambalotti. 2017. "Macroeconomic Nowcasting and Forecasting with Big Data." Staff Report No. 830, Federal Reserve Bank of New York.

Buell, B., R. Cherif, C. Chen, H. Seo, J. Tang, and N. Wendt. 2021. "Impact of COVID-19: Nowcasting and Big Data to Track Economic Activity in Sub-Saharan Africa." IMF Working Papers 21/124, International Monetary Fund, Washington, DC.
Chollett, F., and J.J. Allaire 2018. *Deep Learning with R*. Shelter Island, NY: Manning Press.

Flach, P. 2012. *The Art and Science of Algorithms that Make Sense of Data*. London: Cambridge University Press.

Giannone, D., L. Reichlin, and D. Small. 2008. "Nowcasting: The Real-Time Informational Content of Macroeconomic Data." *Journal of Monetary Economics* 55: 665–76.

Efron, B., T. and Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* London: Cambridge University Press.

International Monetary Fund. 2020. *Regional Economic Outlook: Sub-Saharan Africa—COVID-19: An Unprecedented Threat to Development.* Washington, DC, April.

International Monetary Fund. 2020. *Regional Economic Outlook: Sub-Saharan Africa—A Difficult Road to Recovery.* Washington, DC, October.

International Monetary Fund. 2021. *Regional Economic Outlook: Sub-Saharan Africa—Navigating a Long Pandemic.* Washington, DC, April.

Iyer, T., and A. Gupta. 2019. "Nowcasting Economic Growth in India: The Role of Rainfall." Economics Working Paper Series No. 593, Asian Development Bank, Manila.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2021. *An Introduction to Statistical Learning, Second Edition.* New York: Springer.

Kennedy, P. 2008. *A Guide to Econometrics, 6th Ed.* Cambridge, MA: MIT Press.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. New York: Springer.

Marini, M. 2016. "Nowcasting Annual National Accounts with Quarterly Indicators: An Assessment of Widely Used Benchmarking Methods." IMF Working Paper 16/71, International Monetary Fund, Washington, DC.

Narita, F., and R. Yin. 2018. "In Search of Information: Use of Google Trends' Data to Narrow Information Gaps for Low-income Developing Countries. IMF Working Papers 18/286, International Monetary Fund, Washington, DC.

Richardson, A., T. Mulder, and T. Vehbi. 2021. "Nowcasting GDP using machine-learning algorithms: A real-time assessment." *International Journal of Forecasting* 37: 941–48.

Rhys, H. 2020. *Machine Learning: with R, the Tidyverse, and MLR*, Shelter Island, NY: Manning Press.

Solmaz, S., and M. Taheri Sanjani. 2015. "How External Factors Affect Domestic Economy: Nowcasting an Emerging Market." IMF Working Paper 15/269, International Monetary Fund, Washington, DC.

Tiffin, A. 2016. "Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon." IMF Working Paper 16/56, International Monetary Fund, Washington, DC.

VII. APPENDIX: CONCEPTS AND TOOLS IN MACHINE LEARNING

There is no widely accepted consensus on the definition of machine learning. Broadly, the field has its origins in computational statistics and is chiefly concerned with the use of algorithms to identify patterns within a dataset (Kuhn and Johnson 2016). The actual algorithms can range from the OLS regression to the most complex "deep learning" network; but machine learning is distinguished by its often single-minded focus on predictive performance—indeed, the core of machine learning is the design of experiments to assess how well a model trained on one dataset will predict new data.

As such, machine learning is almost ideally suited to the nowcasting problem, where the goal is to use all currently available information to predict what future GDP releases will say about the current environment. For this purpose, it does not matter whether an indicator is a causal factor that shapes GDP or whether it is instead a symptom of GDP growth. What matters is simply that the indicator contains information about the current state of the economy (Tiffin 2016).

In this regard, the growing popularity of machine-learning techniques stems from their ability to discover complicated patterns that have not been specified in advance. In economics in particular, the world is complex, and everything is connected. Hence, a useful predictive model should ideally be able to sift efficiently through a broad range of potential variables, identifying the relationships, thresholds, and interactions that are most reliably and robustly informative.

The Essence of Machine Learning: Overfitting vs. Underfitting

But the use of complex, flexible models often comes at a cost—they can work too well. Fitting is easy, prediction is hard. And a key danger of using a complex model is that it will almost always fit the existing sample well. Indeed, a sufficiently complex model should be able to fit the data perfectly. But that is no guarantee of future predictive performance, as a perfect fit suggests that the model has not only captured the underlying predictive relationships in the data (the signal) but has also modeled the data's idiosyncratic noise. When making predictions *out of sample*, therefore, such a model is likely to do relatively poorly. (Figure A1). Indeed, "data mining" in traditional econometrics is often disparaged, as it risks producing a model tailored to the peculiarities of a particular dataset, which will consequently be misleading in terms of what it implies about the underlying data-generating process (Kennedy 2008).

Figure A1: Model Complexity and Forecast Error
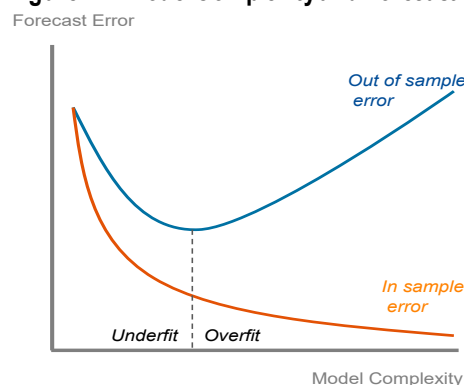
Source: IMF staff.

In machine-learning parlance, this is called the "overfitting" problem. Much of the machine learning literature is tightly focused on addressing this very risk. The key goal of any machine-learning technique is to optimize its likely out-of-sample performance. And using various simulation techniques, such as cross validation, the aim is to filter out noise and produce a model that deemphasizes the peculiarities of a particular dataset. The result, from a machine learning

perspective is the best of all worlds; producing all the potential benefits of a data-mining approach—in terms of capturing key relationships without assuming that the modeler knows exactly how the world works—while avoiding data mining's potential pitfalls.

Key Concepts and Algorithms

Regularization methods include ridge, least absolute shrinkage and selection operator (LASSO), and elastic net and are often called "shrinkage." These methods help to prevent overfitting by stabilizing the parameters of a model and "shrinking" them toward zero. As a technique, regularization can be applied to most machine-learning problems but is most commonly used in linear modeling, where it shrinks the slope parameter of each variable toward zero. The most well-known regularization techniques for linear models are ridge regression, LASSO, and elastic-net regression.

Ridge regression is very similar to OLS, except that the coefficients are estimated by minimizing a slightly adjusted loss function, which imposes a penalty that increases with the squared magnitude of the model's coefficients (the "L2 norm"). More formally, the ridge regression takes the following form:

$$\beta_{ridge} = \text{argmin}\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + Penalty\right],$$

$$Penalty = \lambda \sum_{j=1}^{p} \beta_j^2$$

The main goal of the added term is to discourage model complexity by penalizing large coefficients, effectively biasing them toward zero. The deliberate introduction of a bias may seem counterintuitive at first but recall that the goal is to optimize *out-of-sample* performance, rather than in-sample fit. In cases the in-sample data includes idiosyncratic noise, the addition of some bias will often improve the reliability of out-of-sample predictions (known as the bias-variance tradeoff) (Figure A2).

Figure A2: Bias-Variance Tradeoff
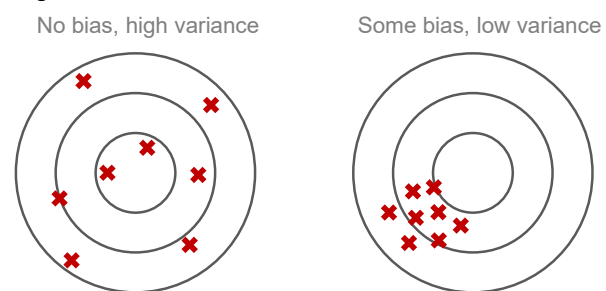
Source: IMF staff.

As a key part of the penalty term, λ is a hyperparameter that determines the term's overall importance—a value of zero for λ would remove the penalty altogether and would be equivalent to fitting an OLS model, whereas increasing the value results in a progressively less complex model, with coefficients squeezed closer and closer to zero. The value for λ is typically defined ex ante and is

usually set to optimize expected out-of-sample performance, as determined by sampling techniques such as cross validation.

LASSO regression is very similar to ridge regression, except that the LASSO uses a penalty that increases with the *absolute value* of the model's coefficients (the "L1 norm").

$$Penalty(\beta_{Lasso}) = \lambda \sum_{j=1}^{p} |\beta_j|$$

Similar to the ridge regression, the penalty discourages complexity by biasing the coefficients toward zero. Unlike the ridge regression, however, with a large enough penalty some of these coefficients will be set *exactly equal* to zero, effectively removing the corresponding variable altogether. Hence, LASSO can be used for feature selection—so if a nowcasting problem has a large number of candidate variables, the model will automatically choose the subset of those variables that are most informative in predicting GDP. Again, the higher the value for λ, the more variables are discarded. And λ is typically chosen via cross validation to optimize expected out-of-sample performance.

Elastic net regression is a compromise between the ridge and LASSO regressions and contains a hybrid of the two penalties. The net effect is that, as the regression's penalty increases, some variables have their coefficients squeezed toward zero, while other less-informative variables are removed from the model entirely.

$$Penalty(\beta_{elastic\ net}) = \lambda \left(\sum_{j=1}^{p} (1-\alpha)\beta_j^2 + (\alpha)|\beta_j| \right)$$

The hybrid penalty is a convex sum of the ridge and LASSO penalties, with relative weights determined by an additional hyperparameter α. Once again, the combination of *α* and *λ* are chosen ex ante and are typically set via cross-validation to optimize expected out-of-sample performance. For further details on these techniques, please see Rhys (2020).

Ensemble methods include decision trees, random forests, gradient boosting, and adaptive boosting.

Decision trees are flowchart-like structures designed to predict a particular output. For instance, imagine a flowchart where each level is a question with a binary yes or no answer (for example, "is a country's debt larger than 60 percent of GDP?"), followed by other levels with binary answers. Following the chart and answering the questions one by one, eventually gives a solution to the initial problem. The challenge is to come up with the right initial questions. To arrive at the solution, the decision tree algorithm repeatedly partitions the predictor space into two sets, starting with an initial split that decreases the prediction error the most: that is, the algorithm considers every possible split on every possible predictor variable, and chooses the one split on the one variable that best separates the sample into the two most dissimilar subsamples (based on the predicted outcome).

These binary partitions then continue until the termination of the tree, and are recursive—that is, each subsequent split only considers the subsample under which it falls, rather than the whole dataset. The result is an efficient set of yes/no questions that quickly sorts the sample into similar bins. The prediction for that bin is then the average outcome of all "similar" observations within that bin.

Decision trees are computationally efficient and work well for problems where there are important non-linearities and interactions. A potential downside, however, is that they are extremely flexible and so are prone to overfitting—with enough questions a large tree can feasibly sort each observation into its own individual bin, fitting the sample perfectly. A common solution, therefore, is to shorten or "prune" the tree by imposing a penalty for an overly long or complex structure. As a regularization technique similar to the penalties outlined above, the ideal degree of complexity is then chosen using cross-validation.

The random forests algorithm takes the decision tree as a basic building block, but instead of pruning the tree, the algorithm filters out any idiosyncratic noise by bootstrapping a large number of separate (unpruned) trees, and then takes the average prediction of those trees. By constructing many trees under similar but randomly drawn conditions, the algorithm minimizes the variance of the model's predictions without necessarily introducing any added bias.

Random forest is one of the most successful general-purpose algorithms currently available. It requires almost no input preparation, since it can handle a range of different predictor types (binary, categorical, numerical) without the need for scaling. It implicitly incorporates an element of feature selection, is quick to train, and can be applied to a wide range of modeling tasks. When the popular machine-learning competition website Kaggle was established in 2010, the random forest algorithm quickly established itself as a platform favorite—at least until 2014, when *gradient boosting machines* took over (Chollett and Allaire 2018).

The *gradient boosting* algorithm is much like a random forest, in that it entails aggregating a large number of decision trees. But rather than averaging over all models at once, boosting is a sequential ensemble algorithm, in which the trees are constructed one at a time, and in which each tree aims to learn from the mistakes of the previous one. For example, the gradient boosting algorithm starts out by training an initial decision tree on the full dataset. Taking the predictions of the first tree, a second tree is then trained to predict the errors from the first. A third tree is then trained to predict any residual errors from the second, and so on. The final prediction is then the sum of the individual predictions from all of the trees.

Adaptive boosting (ADABoost) takes a slightly different approach. Instead of predicting the errors of the previous model, each iteration tries to generate predictions based on a reweighted dataset, where the weights are determined by the previous model. More weight is given to instances that the previous model handled poorly, and less weight is given to those it handled well. The final prediction is a weighted sum of all the models, with weights determined by each model's accuracy. For further details on these techniques, please see Efron and Hastie (2016)

Kernel Methods include support vector regression (SVR). The SVR aims to address the potential role of complex, non-linear relationships by first mapping the data to a higher-dimensional representation where the line of best fit can be expressed as a simple linear hyperplane. There are different possible ways of finding this hyperplane, but in the case of an SVR, it is found not by

minimizing the sum of squared residuals, but instead by minimizing the sum of absolute errors, counting only errors that are greater than a certain threshold, and scaling those errors by a cost parameter.

Mapping data to a higher-dimensional space—where the regression problem becomes simpler—is easier said than done and is often computationally intractable. However, kernel-based methods employ a shortcut (the "kernel trick"). To find a good hyperplane in the new representation space, it is not required to calculate the new coordinates of the data in that space. Instead, it is enough to simply compute the *distance* between any two points in the space. This can be done efficiently using a *kernel function*—a tractable operation that maps any two points in the initial data space to their distance in the new space, without having to actually calculate the new representation space explicitly.

The kernel function is typically chosen ex ante and can take a range of forms (linear, polynomial of various degrees, sigmoid, and so on). The most popular, however, is the Gaussian Radial Basis Function (RBF) kernel:

$$k(x_i, x_j) = \exp\left(\left|x_i - x_j\right|^2\right)$$

Kernel-based methods are one of the few machine-learning approaches that are amenable to mathematical analysis, making them well understood, easily interpretable, and historically popular. But they do not always scale well to large datasets, so their popularity has eased over the past decade or so. For further details on these techniques, see James and others (2021).