Advancing R&D Expenditure Predictions: The Power of Google Trends Data in Nowcasting

Leila Aissa, Aymane Lamyaghri and Abdessalam Derouich
EPFL

***Abstract*—This report presents a new approach to now-casting Research and Development (R&D) expenditure. Using Gross Domestic Expenditure on R&D data, macroeconomic variables, and Google Trends, we develop models to predict R&D expenditure on a yearly basis and a quarterly basis. Best results are achieved using only up-to-date Google Trends data with a MAPE (Mean Absolute Percentage Error) of approximately 2.5% on a yearly basis, and 4.2% on a quarterly basis.**

I. Introduction

Our study consists of predicting Gross Domestic Expenditure on Research and Development (GERD) [4] which encapsulates the total expenditure (both current and capital) on R&D activities are conducted by all resident organizations, including companies, research institutes, universities, and government laboratories within a nation's territory. Up to our knowledge, no previous published research has tackled this task.

First, we focus on predicting the yearly R&D expenditure using historical GERD data and macroeconomic variables. We then follow an innovative approach to perform the same task using only Google Trends data. Finally, we provide a promising technique for estimating R&D expenditure on a finer granularity.

II. R&D estimation on a yearly basis GERD + macroeconomic variables

Certain macroeconomic variables, such as Gross Domestic Product (GDP), inflation rates, and unemployment figures, are traditionally regarded as key indicators of a nation's economic health. This section explores the potential of these variables to predict R&D expenditures. The premise is that a nation's economic strength, mirrored by these variables, might correlate with its investment in research and development. Our study aims to investigate this correlation, assessing the viability of these macroeconomic factors as predictors of R&D expenditure.

A. Data and preprocessing

Our study focuses on 9 countries: Canada, China, France, Germany, Japan, South Korea, Switzerland, the United Kingdom and the United States of America. We collect data available between 2005 and 2020. The lower bound was imposed by the availability of Google Trends (GT) data (not available before 2005), the GT data not used in this section, will be referred to in the next one. The upper bound was set because of the availability of the ground truth values for the macroeconomic variables.

1) GERD data: The dataset contains for each country and each year the 2 previous available R&D expenditure values.

2) Macroeconomic variables: We select relevant macroeconomic variables that could provide a comprehensive overview of a nation's economic state, capturing elements of wealth, economic conditions, and competitiveness such as GDP (Gross Domestic Product), inflation rate, unemployment, etc. We convert data to obtain the same unit per feature across the dataset. This data is sourced from the International Monetary Fund [3].

B. Metrics and Models

This subsection details our approach to evaluating the utility of macroeconomic variables in R&D expenditure prediction. We apply various models to two sets of data: one using GERD data alone and another combining GERD with macroeconomic variables. The goal is to assess the added value, if any, of macroeconomic indicators in this context. In doing so, we also aim to identify the most performant model. From linear regression to advanced neural networks, our analysis will help us discern the effectiveness of each model in capturing R&D spending patterns.

1) Metrics: We use MAE and MAPE as metrics to assess the performance of the models :

MAE (Mean Absolute Error) gives an average error magnitude and is measured in the same unit as the data i.e. percentage of GDP.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAPE (Mean Absolute Percentage Error) provides the error as a percentage, making it unitless and thus easy to interpret and more intuitive to stakeholders.

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

2) Models: Our study utilizes a variety of models to accurately predict and understand R&D expenditure :

- **Linear Regression:** This basic statistical tool is used for its simplicity and interpretability, providing a baseline understanding of the relationship between variables.
- **ARIMA (AutoRegressive Integrated Moving Average):** A robust model for time-series data, ARIMA is adept at capturing and forecasting trends over time, making it ideal for longitudinal data analysis.

- **Neural Network (1-layer):** A single-layer Neural Network is used for its straightforward architecture, serving as a comparative standard to assess the complexity and effectiveness of more advanced models in the study.
- **LSTM (Long Short-Term Memory):** Despite concerns about overfitting, LSTM is explored for its capability in handling sequential data and long-term dependencies, offering a deep learning perspective to the analysis.
- **KNN (k-Nearest Neighbors):** This algorithm is tested using cross-validation with different values of k to refine the model. KNN would be particularly useful for making predictions at a finer granularity, such as quarterly R&D expenditure estimates.

C. Predicting using only GERD data

In this subsection, we will employ the models discussed previously to assess their performance using only historical GERD data.

Model \Metric	**MAPE**	**MAE**
Linear Regression	3.78%	0.09
ARIMA	14.31%	0.49
NN (1 layers)	**3.67%**	**0.093**
LSTM	3.95%	0.099
KNN (with CV)	6.33%	0.17

TABLE I
MODEL PERFORMANCE COMPARISON WHEN USING ONLY HISTORICAL GERD DATA AS FEATURES

Analyzing the data from the table I, it is evident that Neural Networks (NN) outperform other models in predicting R&D expenditures using only historical GERD data. Specifically, the Neural Network with multiple layers demonstrates the best performance, yielding a MAPE of 3.23% and a MAE of 0.12 percentage of GDP. This indicates a higher accuracy and reliability compared to other models, including Linear Regression, ARIMA, and LSTM. We also note that although Linear regression performs similarly, this performance of NN is achieved without any specific optimization of the network's parameters, suggesting an inherent advantage of NN in handling this predictive task.

D. Predicting using GERD data + Macroeconomic variables

In this subsection, we aim to predict R&D expenditures using a combination of historical GERD data and macroeconomic variables. If this approach outperforms predictions based solely on GERD data, it would prove the significance of macroeconomic factors in forecasting R&D investments.

E. Discussion and findings

Analyzing the results in the table II for predictions using both GERD and macroeconomic variables, We observe that the general performance of all models when using this combined data is worse than the one obtained using only GERD data. Therefore the addition of macroeconomic variables does not enhance the predictive capability of the model. Moving forward we will use a Neural Network model without the inclusion of macroeconomic data during training.

Model \Metric	**MAPE**	**MAE**
Linear Regression	5.39%	0.13
ARIMA	21.22%	0.54
NN (1 layers)	11.08%	0.26
LSTM	5.46%	0.13
KNN (with CV)	6.33%	0.17

TABLE II
MODEL PERFORMANCE COMPARISON WHEN USING HISTORICAL GERD DATA + MACROECONOMIC VARIABLES AS FEATURES

Macroeconomic variables ultimately prove to be ineffective for predicting R&D expenditures. Additionally, the availability of GERD data is often limited or delayed. This presents a significant challenge in developing timely and accurate predictive models.

III. R&D ESTIMATION ON A YEARLY BASIS GOOGLE TRENDS DATA

In light of those challenges, we shift our predictive focus to estimating the differences between consecutive R&D expenditure values, an approach that offers significant advantages [2] especially in quarterly analysis. This strategy not only enhances accuracy but also yields immediate insights into the impact of each quarter on annual R&D expenditures leveraging the nuanced data provided by Google Trends.

A. Baseline Model for R&D Prediction

The insights gained from the previous part of our study have led us to choose a neural network model as our baseline using only the GERD data. In order to make it a better we optimize it using **5-fold cross-validation** achieving a MAPE of **2.92%**, this will be a crucial benchmark for our subsequent analysis with Google Trends data.

B. Extracting Google Trends Data

For the extraction of Google Trends data, we employ the 'pytrends' library to interface with Google Trends. Our data collection spans 100 categories related to R&D expenditures, covering the years from 2005 to 2020.

The utilization of categories from Google Trends, as opposed to specific keywords, presented a significant advantage in terms of language harmonization across the 9 countries in our study. This approach mitigated the risk of language-based biases in our data set, ensuring a more uniform and reliable dataset for our predictive analysis.

To retain only pertinent data we conducted a correlation analysis between categories and the R&D expenditure values to only keep the most significant categories. We further refined our selection process through **cross-validation**, identifying the optimal correlation threshold for retaining a category. Details of this cross-validation process and the criteria for category selection will be elaborated in subsequent sections. Here are the top 10 most correlated categories we identified:

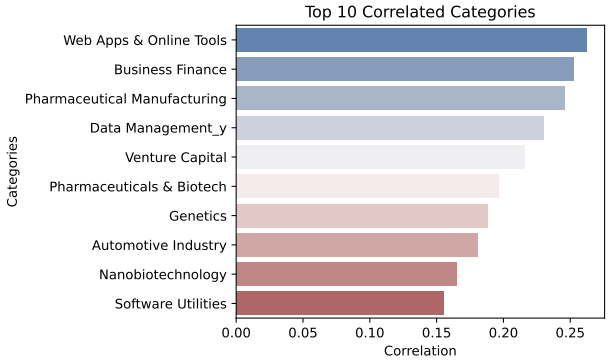

Fig. 1. Top 10 Correlated Categories in Google Trends Data

C. R&D Expenditure Predictions Using Google Trends Data

This section delves into the utilization of Google Trends data for predicting yearly R&D expenditures, highlighting the advancements achieved through this approach.

The initial step involved rigorously fine-tuning our fully connected neural network model to effectively incorporate Google Trends data. Critical parameters such as network width, depth, and hyper-parameters were optimized through a **5-fold cross-validation** process. The details of the model optimization are in the next section.III-D

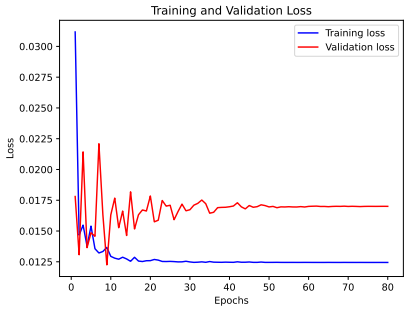

Fig. 2. Training and validation loss over epochs

Metric	Test Set	Training Set
MAPE	2.53%	2.13%
MAE	0.064	0.055

TABLE III
MAPE AND MAE FOR TEST AND TRAINING SETS AVERAGED OVER CROSS-VALIDATION FOLDS

The figure 2 illustrates a rapid decline in the neural network's loss after only a few iterations, which can be attributed to the high number of parameters in the model. Additionally, the table III indicates that the model avoids overfitting as we can see that the error values for the test and training set are close. Furthermore, the model has shown remarkable improvement, achieving a MAPE of **2.53%**, which surpasses the performance of the baseline model. This significant advancement shows the effectiveness of incorporating Google Trends data into predictive modeling for R&D expenditures.

The success of this model not only consolidates its efficacy in yearly predictions but also sets the foundation for extending these predictions to a quarterly basis. The subsequent sections will explore how this approach can provide more granular insights into R&D expenditure trends on a quarterly scale.

D. Optimization of the Neural Network for R&D Prediction

This subsection focuses on the specific optimizations applied to our fully connected neural network, crucial for enhancing the model's predictive performance. It is possible to observe a small variation in the values of the tuned hyper-parameters across different runs, but this does not impact our results and conclusions.

1) Neural Network Width: Adjusting the number of neurons in the neural network significantly impacts its learning and generalization capabilities. We employ cross-validation to ascertain the optimal network width. The results of these evaluations are succinctly depicted in the following figure 3.

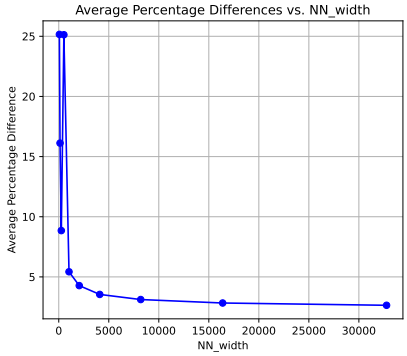

Fig. 3. Optimization of Neural Network Depth

The results indicate that a network width of 2048 neurons is sufficient for achieving near-optimal results. We refrained from further increasing the width to maintain efficiency and allow for additional depth in the network.

2) Neural Network Depth: Optimizing network depth yields the following results :

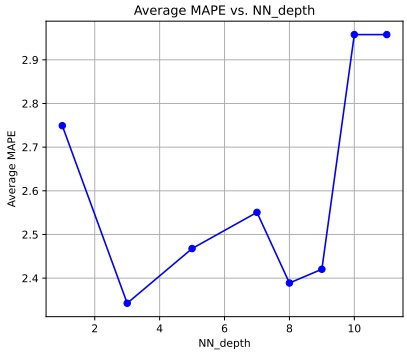

Fig. 4. Optimization of Neural Network Depth

The model exhibits similar performance across depths from 3 to 9, likely due to the optimization's randomness. Consequently, we heuristically selected a depth of 5.

3) Correlation threshold: Optimizing correlation thresholds yields the following results :

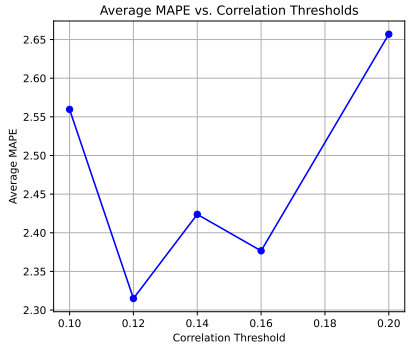

Fig. 5. Optimization of correlation threshold

A correlation threshold of 0.12 represents a reasonable choice, as it corresponds to the lowest point on the graph.

4) Number of epochs to run: Considering different epochs allowed us to observe the following performance figure 6:

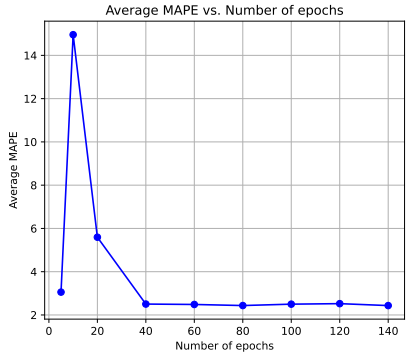

Fig. 6. optimization of number of epochs

This observation aligns with our chosen setting of 80 epochs for the model's training.

5) Batch Size and Learning Rate Adjustments in Fully Connected Neural Network: In our fully connected neural network, batching and learning rate scheduling were key techniques used for optimization. Specifically, we determined that a batch size of **8** was optimal. Additionally, an initial learning rate of **0.001**, coupled with a scheduler that decreases this rate after a certain number of epochs, was found to be most effective. These adjustments were critical in enhancing the overall performance of the model. For a more detailed analysis of the choice of these hyperparameters and their corresponding plots, please refer to the appendix.

IV. R&D Estimation on a Quarterly Basis Google Trends Data

Pivoting our approach to quarterly predictions, we explore the potential of Google Trends data for estimating R&D expenditures within individual quarters. This nuanced analysis is significant due to the absence of quarterly R&D data, necessitating an innovative adaptation of our modeling strategy.

A. Training Strategy for Quarterly Analysis

Faced with the limitation of unavailable quarterly R&D expenditure data, we adapt our model training to suit this finer granularity. The training is conducted in a manner similar to the yearly model, with an important modification: the Google Trends data were aggregated by summing values for each quarter and not averaged as in the yearly model. This approach is designed to intuitively align the model with the annual nature of R&D data.

B. Quarterly Predictions and Model Performance

In the testing phase for quarterly predictions, our approach differs from the yearly model due to the absence of direct quarterly R&D expenditure data. Instead, our model is applied to quarterly Google Trends data, with the predicted values for each quarter summed to estimate the R&D expenditure for the following year. This indirect method has proven effective in capturing the trends and fluctuations inherent in quarterly data. The model achieves an average MAPE of 4.6%.
While this outcome is not as robust as the yearly model, it highlights the model's capacity to deliver insightful predictions, demonstrating its adaptability and effectiveness even without direct quarterly R&D expenditure data.

V. Conclusion

In this study, we demonstrate the efficacy of Google Trends data in enhancing R&D expenditure predictions achieving a MAPE of 2.5% annually and 4.2% quarterly. These results not only validate the effectiveness of alternative data sources in economic forecasting but also pave the way for their application in more granular, time-sensitive analyses. Ultimately, this research highlights the potential of leveraging innovative data sources to improve the precision and dynamism of economic models in today's data-driven landscape.

VI. ML Ethics

When trained on data from 2005 up to 2015 and evaluated exclusively on data from Germany and Great Britain, the model exhibits a commendable accuracy for Germany with an average MAPE of 1.8%. However, this performance significantly deteriorates for Great Britain, where the average MAPE escalates to 7.6%.

The figure 7 represents the prediction of R&D expenditures for both countries between 2016 and 2020

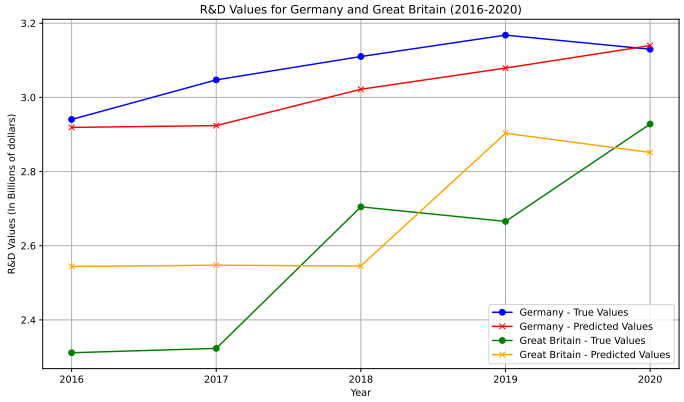

Fig. 7. R&D values for Germany and Britain

This disparity in model performance between Germany and Great Britain may indicate the presence of some ethical concerns :

- **Google trends Reliability:** It may be tempting to generalize this model for more countries but one should take into consideration the political landscape (censorship laws for example) there. The fact that we rely on Google trends data works only under the assumption that the individuals in the analyzed countries use Google for their daily search requests, therefore mirroring their interests, which may discriminate against countries where that is not the case. It is also important to note that in some cases, GT yields different plots when queried for the same search term, period, and country. Such variation should be taken into consideration when using our model or more generally when working with GT [1].
- **Google Dependence:** In the wake of multiple Google data collection concerns ([5]) , making our model rely solely on the results from one monolithic company suspect to ethical mistakes may be detrimental, this however can be solved by diversifying the sources of dynamic "trend data" now that we have proven their important practical role in predicting R&D expenditure.

In conclusion, while the model shows promise in democratizing R&D expenditure globally, one should take into account these ethical aspects when using and interpreting the obtained predictions.

REFERENCES

[1] A. Franzén. Big data, big problems: Why scientists should refrain from using google trends. *Acta Sociologica*, 66(3):343–347, 2023.

[2] T. B. Götz and T. A. Knetsch. Google data in bridge equation models for german gdp. *International Journal of Forecasting*, 35:45–66, 2019.

[3] International Monetary Fund. World economic outlook database. Accessed: 2023-12-21.

[4] Organisation for Economic Co-operation and Development (OECD). *Measuring R&D: Methodologies and procedures*. OECD Publishing, Paris, 2015.

[5] The New York Times. Google privacy settlement, 2022. Accessed: 2023-12-21.

APPENDIX

1) Batch size optimization: Optimizing batch size for our neural network yields the following results:

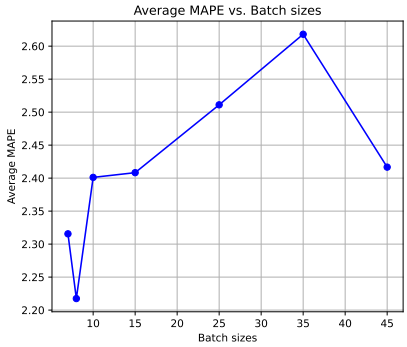

Fig. 8. Optimization of correlation threshold

Our data analysis suggests a batch size of 8 is optimal, balancing computational efficiency and training effectiveness.

2) Neural network initial rate optimization: Optimizing the initial rate for our neural network yields the following results:

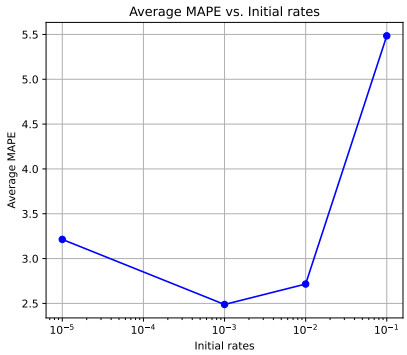

Fig. 9. Optimization of correlation threshold

Our analysis indicates that an initial learning rate of 0.001 is ideal, offering the best balance between convergence speed and stability.