

**Forecasting youth unemployment in the aftermath of the COVID-19 pandemic:  
the Italian case**

**Livio Fenga** <sup>a,\*</sup>,

**Semen Son-Turan** <sup>b</sup>

<sup>a</sup> Italian National Institute of Statistics, Rome, Italy

e-mail: fenga@istat.it

<sup>b</sup> Department of Business Administration, MEF University, Istanbul, Turkey

e-mail: sons@mef.edu.tr

\* Corresponding author

**Abstract**

**Purpose:** This study aims at forecasting NEET unemployment in Italy using a counterfactual scenario, based on an original empirical model, whereby the effects of the COVID-19 pandemic on the NEET rate are factored in and left out.

**Methodology:** An artificial neural network (ANN) model of the type feed-forward, with a Google Trends-generated variable that represents potentially relevant search queries, is employed to backcast, nowcast and forecast Italian NEET unemployment for 2019, 2020, 2021, respectively.

**Findings:** Findings suggest that the Italian NEET unemployment rate will slightly increase in a less than proportional way, absorbing the COVID-19 pandemic's effects in a relatively short time period.

**Research Implications/ Limitation:** Several limitations with respect to the limited sample size and the few number of explanatory variables are remedied through the use of an adequate methodology.

**Originality:** The use of an ANN in youth unemployment studies during a pandemic of the present scale is, to the best of the authors' knowledge, unprecedented.

**Keywords:** Artificial neural network, COVID–19, youth unemployment, pandemic, maximum entropy bootstrap

## I. Introduction

*“Life is what happens to you while you are busy making other plans”*

- *John Lennon*

...and just as the world had come to grips with the fact that combating inequality in a concerted effort is a *conditio sine qua non* for sustaining our livelihoods, the novel coronavirus (COVID-19) disease broke out and swept across the globe creating a crisis that has by no means “left no one behind” (Agenda 2030, 2015).

Recognized as “an unprecedented humanitarian crisis since the Second World War” with far-reaching immediate repercussions that have pushed numerous health care systems to their limits, the COVID-19 pandemic not only strains national economies, but also creates a new *Weltschmerz* that threatens to challenge and erode the “meagre” advances (SDG Progress Report, 2019) made thus far on the United Nations (UN) Sustainable Development Goals (SDGs).

Indeed, the gap between the rich and the poor segments of society, the developed and developing nations, is destined to widen due to extended lockdowns, layoffs, furloughs, salary reductions, shutdowns and ensuing bankruptcies. The labour market is already at the verge of a breakdown with 68% of the world's workforce starkly impacted (ILO Monitor, 2020). According to forecasts one of the strongest economies, the US, will suffer a 20 million

potential decline in the number of employed employers (Coibion *et al.*, 2020).

This study concentrates on Italy, the epicenter of the COVID-19 pandemic, with nearly 200,000 confirmed cases and a dead toll around 27,000. The country's economy already suffered dramatic impacts due to the 2008 financial crisis and has, till then, not been able to recover. This is evidenced by a sluggish growth in real GDP of 0.8% and 0.2% in 2018 and 2019, respectively, accompanied by subdued domestic demand and real disposable income remaining below pre-crisis levels (EC Country Report, 2020).

As one of the worst-hit countries has started to gradually ease its lockdown that had started on March 9, 2020, it remains unknown how the effects of COVID-19-related potentially severe demand and supply-side shocks on its already fragile economy will ripple through one of the most vulnerable groups in the Italian labor market, one that may never have gotten around to living *la dolce vita* (Fellini, 1960): young people neither in employment nor in education and training (NEET). The NEET rate conveys the proportion of youth (aged 15-29 years) who are outside the educational system, not in training and not in employment (Please see Appendix I for a detailed calculation), and thus serves as a broader measure of potential youth labor market entrants than youth unemployment (ILOSTAT, 2020).

Although contested in its definition, specifically due to the heterogeneity of young people included in this statistic (Quintano *et al.*, 2018), the NEET rate has become significant enough to be addressed as an indicator of youth employment in the Agenda 2030 for Sustainable Development, whereby UN member states committed to increase youth employment opportunities and to substantially reduce the proportion of youth not in employment, education or training (SDG 8.6) (Agenda 2030, 2015).

NEET unemployment is a critical issue across many parts of the world with common factors that potentially hinder their entrance into the labour market, such as the lack of general

competencies or job-specific skills (Demidova and Signorelli, 2012), or the lack of jobs that demand entry level skills (Balan, 2015). Furthermore, especially during economic crises, high levels of youth unemployment dramatically increase the share of discouraged young people who give up job searching altogether (De Luca, 2019).

In Italy, especially in the south, being NEET has multiple and often intertwined causes, furthermore in Italy, the segmentation across territory is very marked, and job opportunities are unevenly distributed among the labour force, with the marginalization of specific segments of the working population (De Luca, 2019). Additionally, Italian families tend to support their children into adulthood and thus may contribute to the delay of their access to the labour market (Di Stefano, 2019).

With the economic shocks caused by COVID-19, global unemployment levels are already estimated to increase between 5.3 and 24.7 million, however projections show that this effect will be aggravated sharply for youth employment since youth (15-24) are already 3 times more likely to be unemployed than adults (UN, 2020).

In the fourth quarter of 2019, Italy's NEET (17.8%) rate compared to the Euro area average of 10.0% was the highest among all EU countries (Eurostat, 2020). With the country's labor market already in a precarious situation before the pandemic, the NEET rate, post-peak has increased to a hefty 28% in March 2020 (compared to an unemployment rate of 8.4%) (Eurostat, 2020).

With the assumption that the country could suffer irreversible damages to its economy in case of prolonged recession or a second wave of infections, the evolution of the Italian NEET population deserves particular investigation.

To this end, the present study sets out to develop a forecasting model that provides insight into the development of the Italian NEET rate over the post-COVID-19 period.

Different from existing models that employ “traditional” (youth) unemployment indicators found in literature such as GDP growth, inflation, productivity, foreign direct investment, consumer sentiment (Choudhry *et al.*, 2012; Demidova and Signorelli, 2012; Bruno *et al.*, 2014; Quintano *et al.*, 2018), among others, which recently also tend to encompass new Google Trends (GT) variables (Naccarato *et al.*, 2018), this study presents a novel approach to youth unemployment forecasting: the use of an artificial intelligence approach of the type artificial neural networks (ANN) that factors in COVID-19 effects.

The findings contribute to the scarce literature on youth unemployment during economic crises while differentiating itself from the latter due to the wider reach of implications of the COVID-19 pandemic on the society, the economy and political systems at large.

The remainder of this paper is structured as follows: The following section (II) provides a succinct literature review on studies on Italy focusing on explaining and modelling youth unemployment followed by the methodology section (III) that presents the empirical model. Subsequent to the analysis of findings (section IV), the discussion (section V) is presented. The paper concludes with suggestions for potential further avenues of research (section VI).

## **II. Literature Review**

This section provides a short overview of contextually-relevant youth unemployment studies. While the literature on unemployment is relatively broad, studies on youth unemployment and especially, during crisis periods, is very scarce.

Most recently, a few empirical studies on youth unemployment in Europe during crisis periods have emerged, which may be explained by the stark rise in youth unemployment after the recent economic crisis, and particularly, in high proportions relative to the unemployment rate.

Choudhry and Marelli (2012) using Arellano-Bond dynamic panel estimation of ca. 70 countries (including Italy) between the years 1980-2005, determine that different types of financial crises have a significant impact on youth unemployment. This effect is said to go beyond the impact caused by changes in GDP and its lagged version while the persistence of adverse effects last five years after inception of the crisis.

Bruno *et al.* (2014), employ Generalized Methods of Moments (GMM) and bias-corrected Least Squares Dummy Variables (LSDV) dynamic panel data estimators to assess the impact of the 2008 financial crisis on the NEET rate and the youth unemployment rate in EU regions using Eurostat data for the 2000-2010 period. Their findings suggest that NEET rates are persistent.

De Luca *et al.* (2019) through an Autoregressive Distributed Lag (ARDL) model with a seasonal frequency, attempt to analyse the changes occurred in the share of Italian NEETs (whereby the authors' definition of NEET is broader than the present paper's and includes NEET aged 15-29) from 2005 to 2016. Their findings suggest that in addition to the highest share of NEETs, high percentages of early school leavers and low levels of highly educated youth, Italy also shows the highest inactivity rates compared to other countries such as Greece and Spain.

Following the seminal work of Ginsberg *et al.*, (2009), research on unemployment using GT information (Choi and Varian, 2009; Askitas and Zimmerman, 2009; D'Amuri, 2009) through the use of online search engines, and particularly the world-wide leading search engine Google, has become widespread. This may particularly be attributed to the fact that Google data is available free-of-charge, allows the construction of real-time aggregated data with high frequencies on the search for those keywords of interest to the researcher which receive a significant amount of traffic (Askitas and Zimmermann, 2009), and official

economic statistics are published with a lag (Doerr and Gambacorta, 2020).

GT-based unemployment models are shown to outperform most of the others, with their relative performances improving with the forecast horizon, (D'Amuri and Marcucci, 2017) and their increased predictive accuracy (D'Amuri and Marcucci, 2010; Anvik and Gjelstad, 2010; Vicente *et al.*, 2015), especially when the forecasting period is short (Nagao *et al.*, 2019). Such studies employ various econometric methods integrating GT data as explanatory variables alone, or together with traditional official unemployment statistics.

D'Amuri (2009) determined that GT-based models perform better in predicting unemployment in Italy than the other ones with lower mean squared errors (MSE) at multiple estimation lengths and forecast horizons. The author concludes that even in countries where only quarterly unemployment rate data are available and the GT generated time series is short, this variable should still routinely be included in models forecasting unemployment developments.

As mentioned in the beginning of this review, given that the youth unemployment literature by itself is scarce, there are even fewer studies on youth unemployment employing GT data.

Using a modified version of the Kalman filter taking into account the twofold issue of non-stationarity and multiple frequencies of their data, Fondeur and Karamé (2013) demonstrate that the inclusion of GT data in their model improves unemployment predictions on youth unemployment in France, compared to a competing model that does not employ such.

Similarly, Naccarto *et al.* (2018) focus on youth unemployment in Italy through the use two different models; an ARIMA model integrating only official youth unemployment data from Italian National Institute of Statistics and a VAR model combining the former series with the Google Trends query share. Their results show that the use of GT information leads to an average decrease in the forecast error.

### **III. Methodology**

#### **1. Feedforward neural network with autoregressive inputs**

Often, the choice of the “right” model for the problem at hand envisions the construction of a set of competing models (the competition set) – preferably belonging to different classes of methods – out of which the best option is selected.

As in the case of this paper, frequently, such a choice is made according to a data-driven strategy. Such an approach is justified by two main factors: the limited sample size and the overall pattern followed by the data. The former is a strong constraint ruling out a big chunk of statistical parametric models, especially the non-linear ones, whose generally complex structures require “big enough” information sets. On the other hand, by visually inspecting the specific pattern expressed by the data, it is clear how pure linear structures are not suitable for the job.

Consequently, the competitions set has been built with the following statistical methods: ARIMA (Autoregressive integrated Moving Average) (Box *et al.*, 2015), BATS (Box-Cox transformation, ARIMA errors, trend and seasonal components) with no seasonal components (De Livera *et al.*, 2011), AR-ANN (Autoregressive Artificial Neural Network) (Fine, 2006).

Even though belonging to the class of linear models, ARMA has been selected for its remarkable performances obtainable under difficult conditions (i.e. such as non-stationary time series) and its ability to perform adequately when small to moderate non-linear structures are embedded in the data. Even though ARIMA and BATS models in some instances are virtually interchangeable (De Livera *et al.*, 2011), their use is here justified by the fact that many of their configurations are not. For example, the BATS models are non-stationary, while some ARIMA models are.



However, this class of model exhibits a non-trivial drawback: the inability to process anything but the series of interest, and therefore external regressors (either in the form of time series or dummy variables) fall outside its range of applicability.

In this paper an artificial intelligence approach has been chosen. In particular, we employed an artificial neural network (ANN) that presents a feed-forward (FFWD) architecture associated with a signal processing algorithm of the type autoregressive (AR).

## **2. The neural model adopted**

At its core, an ANN is an algorithm designed to reproduce the working nature of a biological brain to the end of extracting from an information set made up of a series of inputs and outputs, its inner hidden patterns (Haykin, 1994). As in the human brain, where a huge number of processing units (the neurons) are interconnected in a hierarchical fashion, ANNs mimic such a structure through an architecture of the type multi-layers.

Typically, an ANN is equipped with an input layer, in charge of picking up the signal, one or more hidden layers, i.e. where the relevant interconnections between the neurons are created, and an output layer, through which the outcome of the net is released. The artificial neurons are the connection units among the different layers and within the layers. In more details, when a connection is established, the net tries to assign a numeric value (called weight) to each of them as a result of a training process. The final set of weights and connections represents the “best” way, for a given prespecified architecture, to process information and yield the outcome in the form of predictions, as in our case, or classification (neural classifiers).

The adopted network is of the type feed forward (FFWD), i.e. the outputs of the input layer, in the form of a linear combination of the weights associated to the inputs and the inputs themselves, are the inputs for the hidden layer. Its outcomes are then modified according to a

nonlinear (activation) function – in our case of the type sigmoid – and sent to the output layer which has only one node representing the predicted value.

The AR-ANN can be formulated as follows:

$$(1) x_t = \alpha_0 + \sum_{i=1}^n x_{t-i} + \sum_{j=1}^h \Psi (\gamma_{0j} + \sum_{i=1}^n \gamma_{ij} x_{t-i}) \beta_j + \varepsilon_t.$$

However, in the present paper, we have included the shortcut connection (stochastic part) in the form of an autoregressive model of order  $p$ ,  $AR(p)$  so that the neural part is employed to capture the residual non-linearities, i.e.

$$(2) x_t = \alpha_0 + \sum_{i=1}^n \alpha_j x_{t-i} + \sum_{j=1}^h \Psi (\gamma_{0j} + \sum_{i=1}^n \gamma_{ij} x_{t-i}) \beta_j + \varepsilon_t.$$

The nonlinear part contains  $h$  hidden neurons, which transform the input variables, weighted by parameters  $\gamma_{ij}$  plus a bias  $\gamma_{0j}$  through the activation function  $\Psi(\cdot)$ , via a nonlinear activation function. Formally, a generic, neuron in the hidden layer takes the form:

$$(3) \Psi (\gamma_{0j} + \sum_{i=1}^n \gamma_{ij} x_{t-i}),$$

where, with the lower case letters  $i$  and  $j$  the number of lags and of hidden neurons are respectively denoted. Each hidden neuron is weighted by a parameter  $j$  before being passed to the output layer.

The training process of our network has been done by minimizing a loss function of the type MSE achieved in an iteratively fashion by finding the best combination of connection weights and biases. To do so, a gradient-based algorithms has been applied to the loss function so that, once a minimum is reached at a given iteration, the error signals is back propagated to tune the weights.

The connections in our net have been optimized through an algorithm of the type gradient descent (with initial weights randomly selected), which iteratively and recursively

changes the weights and biases to minimize a suitable cost function, say,  $\mu$ , which, in our case, is the MSE.

In essence, this procedure works by taking the derivative of  $\mu$  with respect to the network parameters at a specific position of  $\mu$ , and updates the parameters in the direction of the negative gradient (down along the loss function). The neural parameters that are supposed to be stored in a vector called  $\omega$  is iteratively adjusted by taking steps proportional to the negative of the gradient. By denoting those steps with the Latin letter  $K$  we have that

$$(4) \omega(K + 1) = \omega(K) - \Delta(\omega(K))$$

$$(5) \Delta_{\omega} = \lambda \frac{\partial \mu}{\partial \omega},$$

with,  $\lambda$  being the network learning rate, which controls at which speed our network learns from the world (how big the steps are taken along the gradient) and  $\frac{\partial \mu}{\partial \omega}$  the gradient.

### **3. The ANN-AR structure determination and model selection strategy**

As explained thoroughly in Anders and Korn (1996) artificial neural networks are very well suited to describe and model nonlinear relationships since they lend themselves to accurately approximate any measurable function to an arbitrary degree. However, it is this inherent flexibility that makes it difficult to specify an adequate neural network based solely on network theory. The authors therefore propose several statistical approaches to determine the right neural network model. The present paper follows one of them; the Network Information Criterion (Murata *et al.*, 1994) which is a generalization of the well-known Akaike Information Criterion (AIC) (Akaike, 1973, 1974). While both of these information criteria have their shortcomings, i.e. their main limitation being the assumption of the asymptotic

normality of the maximum likelihood estimators which renders them irrelevant in case of overparameterized networks (Anders and Korn, 1996), the present model keeps the hidden neurons causing this overparameterization, at a minimum.

Network Information Criterion (NIC) appeared to be a natural choice, given the small sample size available. In fact, AIC even though it is computed using the whole sample is not really an in sample measure. From an asymptotic point of view, in fact, its minimization is equivalent to minimizing the out-of-sample one-step forecast MSE. For neural networks NIC function is defined as:

$$(6) -\frac{2}{T} \ln L(\mathbf{w}) + 2K/T,$$

where,  $K$  being the number of neural weights plus bias. The applied procedure, called MNICE (Minimum NIC Expectation), selects the network 276 competition set, which exhibits the smaller NIC value.

#### **4. The employed data set**

Two sets of data, freely and publicly available, are used in this study: the NEET data for Italy, which are part of the official youth unemployment data, collected and regularly updated by the Italian National institute of Statistics (<http://dati.istat.it/>) and the GT (<https://trends.google.com/trends/>) data for the variable “*door-to-door*”. The yearly NEET time series is depicted in Figure 1 (upper panel) for all the available years, i.e. from 2004 to 2019 whereas the GT data, for consistency, have been aggregated yearly but this time the time span is from 2004 to 2018 for NEET data and 2004 to 2019 for GT data. The related time series is reported in Figure 1 (lower panel). For the sake of the present study, the GT data have been downloaded on 04/23/2020 at 12.02 PM GMT.

Low frequency (annual) data has been used to obtain a series as smooth as possible thereby avoiding the use of mathematical filtering. This enables to keep at a minimum the introduction of artifacts resulting from the application of a statistical model, which, by definition, generates artificial data.

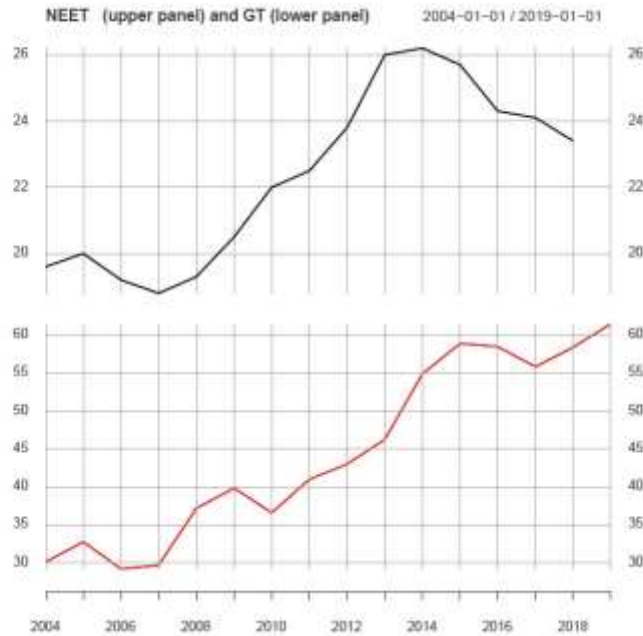
As it is well known, GT data are far from perfection. The two major drawbacks are related to the non-constant population size supporting each data point and to the significant amount of noisy components embedded. Those are mainly related to the inability of Google to discriminate pertinent queries from those unrelated to the labor market. In general, such issues translate into the observed time series in several forms, such as, inconsistent patterns, outliers, and irregularities, which can mask the underlying signal. However, in our particular cases, as it will be discussed later, to model these data we employed a simple regression scheme which, if on the one hand cannot (by design) capture the non-linear components present in the data, and on the other hand, because of that, is able to filter out not negligible portions of uncertainty embedded in the data.

Regarding the NEET data, it should be also said that upward / downward biasing components can affect those data as well. As explained in Quintano *et al.*, (2018) the actual share of inactive individuals comprising the NEET could be underestimated or overestimated due to reasons such as social acceptability or illegal employment. Furthermore, young people who could at “normal” times be categorized as unemployed (due to lay off by their employers) may still be regarded as working since they are receiving support from their local government.

However, the model choice made plays a significant role in minimizing the effects of such distortions. As it will be illustrated, we use an ANN model, which, as a universal function approximator, is designed to extract and project the main features into the futures. Finally,

idiosyncratic components minimization has been facilitated by the low frequency (yearly) at which the whole analysis has been carried out. This particular GT variable has been selected out of a competition set made up with time series which exhibited two features: “a good” level of correlation with the NEET variable and a response to the impulse induced by the 2008 crash. While the former has been assessed through the Pearson correlation coefficient the latter has been evaluated on a visual inspection basis. This criterion has also driven the selection process of the final series along with a function of the type empirical cross correlation, which tries to identify the connection between two correlated sequences (Derrick and Thomas, 2004). However, given the small sample size, this analysis has been carried out taking also into account the spikes within the asymptotic confidence bands. By inspecting both panels of Figure 1, it is clear how the two dynamics are correlated and that, at least for the financial crisis, the GT series shows a certain amount of predictive power.

At this point, one might be wondering why we built two different estimates, when we are clearly interested in the predictions of the World the way it is, i.e. with the COVID-19. This is true, but we believe in the fact that the analysis can be considered completed when both the situations are modeled and thus measurable. By doing so, in fact, we can assess the (future) impact of COVID-19 by considering the area between the two curves as represented in Figure 1.



**Figure 1. NEETs time series (upper panel) from 2004 to 2018 and GT (lower panel) time series from 2004 to 2019**

### 5. Modeling and predicting the GT data

As an external regressor the GT time series related to the keyword “*door-to-door*” (original search term in Italian is “*porta a porta*”, which is typically used by young people to search for low-level jobs that require no specialization), depicted in the lower panel of Figure 1 along with the original series reported in the upper panel, has been employed. This GT series has been chosen for its empirical level of correlation ( $\rho = .84$ ) and for being logically related to the variable of interest.

A simple linear model has been adopted to model (see equation 7) and forecast (see equation 8) variables  $y_t$  and  $x_t$ , respectively:

$$(7) \quad y_t = \alpha + \beta_t + \varepsilon_t,$$

where  $t = 1, 2, 3, \dots, T$

$$(8) y_{T+k} = \alpha + \beta_{T+K},$$

where  $k = 1, 2, 3, \dots, K$  are the steps into the future the model is projected to.

## 6. The employed ANN-AR model

$NNAR(p,k)$  is the short way to refer to our feedforward neural networks with a single hidden layer and lagged inputs being,  $p$  the embedding dimension and  $k$  the number of neurons belonging to the hidden layer.  $NNARx(p,q)$  is its extension designed to encompass external regressors.

### 6.1. The computation of the confidence intervals

The computation of the confidence intervals (CI) has been conducted by means of a bootstrap method of the type Maximum Entropy (Vinod, 2013; Vinod *et al.*, 2009). Even if the procedure is pretty straightforward when the NEET time series is treated in an univariate fashion (no external regressors) it becomes more complicated when the predicted values of the external regressor enter the neural model. This problem is clearly related to the predictions generated for the years 2021– 2022 (being the datum related to the year 2020 known).

Both the forecasting procedure and the computation of the CIs are bootstrap-based. In particular, the latter is done according to the percentile method: in essence we draw a large number of bootstrapped samples using the Maximum Entropy bootstrap so that, by computing the mean of each sample, we are able to create the sampling distribution of the mean. Once a pre-specified level of confidence  $\alpha$  is chosen, we then take the  $\alpha/2$  e  $(1 - \alpha/2)$  percentiles and these are the confidence limits.

### 6.2. Forecasting procedure

For the sake of a better comprehension of what follows, we denote by the symbol  $x_t$  the time



series of interest (NEET), whereas the number of bootstrap replications and the GT regressor are respectively called  $B$  and  $r_t$ , where  $t = 1, 2, \dots, T$  are the time points.

The whole forecasting procedure without external regressor is summarized below in a step by step fashion:

- 1) The NEET time series, say  $x_t$ , is bootstrapped so that a bootstrap matrix  $X_{t,j}^*; j=1,2,3,\dots,B$  of dimension  $T \times B$  is generated. Here, each column represent a “bona fide” replication of the original time series  $x_t$ , and will be referred to with the symbol  $x_{t,j}^*$ ;
- 2) the neural network  $NNAR(p_0, q_0)$  is determined according to minimum NIC procedure and MNICE trained based on the original time series  $x_t$ ;
- 3)  $NNAR(p_0, q_0)$  is applied to each and every bootstrap replication, i.e.  $NNAR(p_0, q_0) [X_{t,j}^*]$
- 4) the forecasted values resulting by applying the neural model are stored in a matrix  $F_{h,j}^*$ , where  $h = 1, 2, \dots, H$  is the forecast horizon;
- 5) the expected value  $\mathbf{E}(\cdot)$  is computed column-wise,  $\mathbf{E}F_{h,j}^*$  for each forecasting horizon  $h$  so that a vector of length  $h$  of predicted values is generated. We call this vector  $p_h^*$ ;
- 6) the percentile method is applied to  $F_{h,j}^*$  so that the CIs are computed.

The forecasting procedure with the external regressor  $r_t$  has been carried out separately for the year 2019 and for the years 2020–2021. In fact, while in the first case the 2019 datum has not been estimated because it is available, in the latter one we need to account for the uncertainty associated with the estimation procedure of those future values.

Therefore, for the estimation of the year 2019 we need to replace step 2 with step A in the list below whereas for the predictions of the years 2020–2021 the above steps 2 and 3 are to be replaced with both the points A and B in the list below:

A. the neural network  $NNARX(p_0, q_0, r_0)$  is determined according to the gradient descent procedure, above briefly recalled, based on the original time series  $x_t$ ;

B. the neural model, with embedded the forecasting equation of equation 9 is applied to each and every bootstrap replication, i.e  $NNARX(p_0, q_0, r_0) [X_{t,j}^*]$ .

**Table 1: Linear model for GT variable**

	Intercept	Std. Error	T value
Intercept	24.4	1.76	13.8
Trend	2.3	.18	13.0

Adjusted R-squared = 0.92

The linear model applied to the GT series is robust, as it can be noticed by Table 1, and thus reliable short term predictions can be reasonably expected.

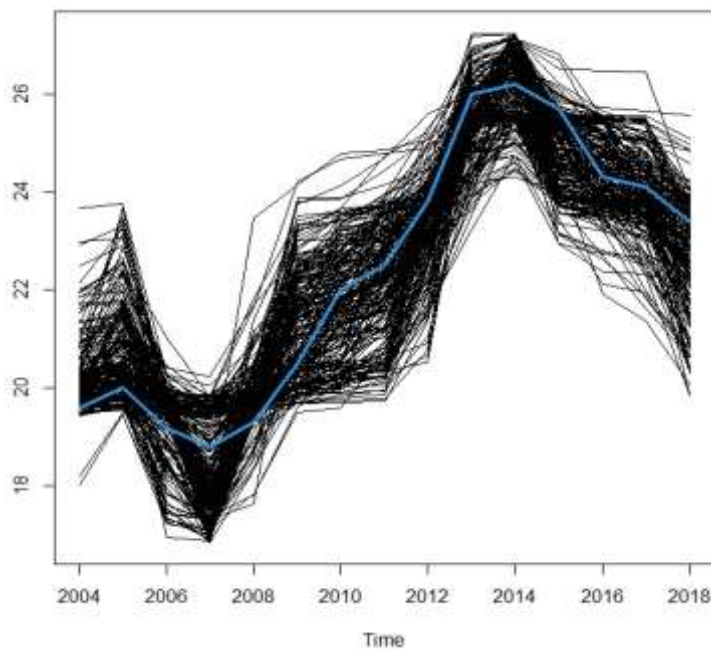
Both the models have been trained using the gradient descent method with derivatives computed on rescaled data, conditioned to a transformation of the type Box Cox. The optimal  $\lambda$  parameter of this transformation has been found through a grid search as the minimizer of the Mean Square Error. In the present case for both the models  $\lambda = -0.48$ , which is very close to  $-.05$ . Therefore the optimal transformation is  $\approx \frac{1}{\sqrt{x}}$ .

The optimal decay ratio has been found using a trial and error approach and set to 0.01. Both networks reach a global convergence (in both non-bootstrap and bootstrap world) in a reasonable computational time. In the case at hand, convergence on 251 time series is reached in about 12 minutes (network without external regressor) and 18 minutes (network with the external regressor).

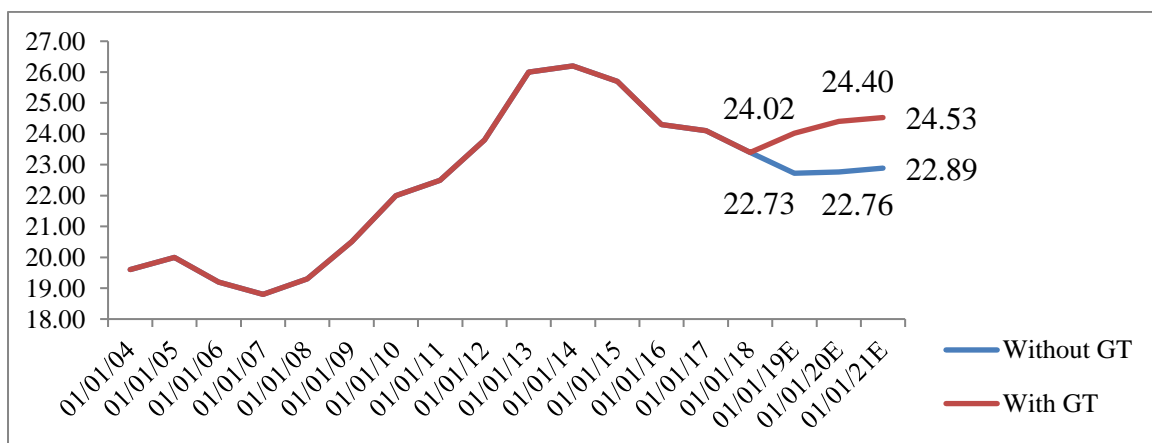
#### IV. Neural models chosen and empirical findings

Following the MNICE procedure, the two models chosen are  $ANNAR(1,1)$  and

$ANNARX(1,2,r)$ , with NIC's of -1.99 and -2.14, respectively. Once the neural models are defined B=250 bootstrap replications are generated so that the matrix of predictions,  $F_{h,j}^*$  of dimension  $h=3 \times 250$  is available. The graphical representation of the 250 pseudoserries along with the original NEET series is demonstrated in Figure 1.



**Figure 2. Original NEET series (in blue) and its B=250 bootstrap replications used in the empirical section**



**Figure 3.** Actual data until 2018 and bootstrapped estimates with and without GT Regressor for the years 2019, 2020, and 2021

Figure 3. demonstrates the bootstrapped estimates with and without the GT regressor, respectively. These findings demonstrate the usefulness of employing GT data in the statistical model. In fact, by introducing into the analysis a signal (our GT) conveying a cross-sectional (i.e. "pervasive") economic "stress", one can get a clear picture of what will happen as opposed to what would have happened in a non-COVID situation. In this sense, the model takes advantage of the information conveyed through the GT variable and, in essence, is able to capture phenomena able to affect, the whole economic system as opposed to just a single sector.

## **V. Discussion**

This study modeled and forecasted NEET unemployment in Italy using a counterfactual scenario, based on an ANN model, whereby the effects of the COVID-19 pandemic on the youth unemployment rate are factored in and left out.

Our findings suggest that the Italian NEET unemployment rate will slightly increase in a less than proportional way which is somewhat consistent with these kinds of shocks. Thus, we can expect that these effects of the pandemic shock are absorbed within a year. Furthermore, in comparison to the 2008 financial crisis, which was structural (endogenous) and purely economic, the COVID-19 crisis is exogenous and has infused across all areas of life. Therefore, it is reasonable to think of this latter effect as being relatively short-term (i.e. lasting for a shorter time). This argument is also supported by the model's results. Indeed, the after-effects following the outbreak of the 2008 financial crisis have lasted for more than five years.

Given that the two crises are therefore very different with respect to magnitude and coverage,

the 2008 financial crisis is not an ideal benchmark. That said, it provides the closest basis for comparison. Different from the post-2008 period though, what is commonly agreed on, is that a drastically new world awaits the COVID-19 survivors with even fewer jobs available.

As with every research, there are potential limitation in this study as addressed and explained below:

In essence, using a model of the class ARIMA would almost surely lead to suboptimal results: The first hurdle is represented by the loss of precious degree of freedom, which would be necessary to achieve stationarity. In fact, the underlying stochastic process seems to realize a hyperbolic trend, which requires an integrated constant of order 3. All the more so, autoregressive, moving average or drift parameters, if detected, must be estimated determining an additional loss of degree of freedom. On the contrary, ANNs are non-linear in nature and the usual restrictions imposed by the Gauss Markov Theorem do not apply. In addition, when properly built and trained, ANN models can learn any type of pattern and provide “accurate” forecast of the future values of the data. In the case of our NEET time series, any attempt to apply an ARMA structure under an integration constant of order 3 resulted in a model of the type ARMA (0,0) that is unsatisfactory.

The usage of a GT variable in this study is justified since its search engine covers almost 94% of all the searches in the country for an extended period of time (Statista, 2020).

The main contribution of this research is the empirical presentation of a counterfactual argument, through the use of a targeted GT variable (“*door-to-door*”), whose inclusion (exclusion) determines the creation of two future scenarios. Both of them are stored in two different “places” according to whether the GT variable enters the statistical model or not. In the latter case, we have a future counterfactual scenario, which, according to the model, will never happen. In essence, it refers to a purely theoretical assumption that the intervention (in

the present case the economic shock caused by COVID-19), which actually happened, never did. By doing so, we built the so-called *baseline* towards which, the effect of what in reality happened (the COVID-19) can be measured in advance. Therefore, this scenario is stored in a hypothetical future which is supposed to not realize itself. Conversely, by including the GT variable into the analysis, we try to extract the underlying stochastic process which, in its future, encompassed the COVID-19 effect. At this point, one might be wondering why we built two different estimates, when we are clearly interested in the predictions of the world the way it is, i.e. with the COVID-19. This is true, but we believe in the fact that the analysis can be considered completed when both the situations are modeled and thus measurable. By doing so, in fact, we can assess the (future) impact of COVID by considering the area between the two curves as represented in Figure 2.

Since the Italian economy has never experienced a shut-down of this scale related to COVID-19, the validity of the model may come into question. To that end, the authors have contemplated designing an experiment to validate the model using the data surrounding the 2008 economic crisis using a training set (e.g 75% of data) and a test set, e.g 25% of the data. However, due to the small sample size such as validation would prove to be severely biased. Nonetheless, our use of the NIC, by nature, serves as a way of validation. Even though this criterion is computed by using the whole sample, it is not really an in-sample measure. From an asymptotic point of view, in fact, its minimization is equivalent to minimizing the out-of sample one-step forecast MSE.

Furthermore, the GT variable with the well-chosen keyword “*door-to-door*” is able to embody an effect that is very similar to the shut-down with respect to magnitude and consequences. As such, Figure 3 demonstrates that the GT variable starts reacting (i.e. increasing) in 2007, that is before the 2008 crisis. This suggests that it is capable to even anticipate the effect of the 2008 economic crisis. By virtue of that, In a sense our GT variable

is able to generate a counterfactual scenario.

As opposed to studies, which use GT variables as complementary explanatory variables next to traditional ones, the only explanatory variable used in this study is the GT variable. Limiting the explanatory variable to one is justified for two reasons: the very limited sample size available and the aim of minimizing potentially correlated predictors risk (e.g. with the purpose of avoiding multicollinearity issues frequently encountered in regression-like schemes). As for the former, the data available, at their best, can support just one source of external information. Regarding the latter point, it has to be said that, the problem at hand is more suitable for external explanatory variables conveying a relevant amount of exogeneity. For example, the use of traditional youth unemployment indicators (as mentioned in the literature review) would inject into the statistical model endogenous components resulting in correlation structures with the predicted variable. In such a situation, in general the explanatory power of the statistical models entertained could be severely compromised.

## **VI. Conclusion**

The current COVID-19 pandemic has underscored how the capability of nations to swiftly, but meticulously, analyze scientific data, anticipate future events, and devise appropriate solutions is essential in devising effective coping mechanisms. In this respect, while the study sample is drawn from Italy, the scientific method employed is geared towards understanding a globally challenging phenomenon (youth unemployment), and has the potential to benefit policy makers from any country. As a result, adequate economic measures and more sustainable stimulus plans can be extended that may significantly mitigate the negative effects on the more vulnerable NEET groups to adverse labour market outcomes.

As mentioned above, it appears that our GT variable exhibits some prediction capabilities, noticeable by looking at the slope change for the year 2007, i.e. the year before the 2008

crash. Whilst such a feature cannot be exploited directly in the present setup, COVID-19 did not give any pre-shock hint to the economy and thus went undetected by our GT regressor, this variable can be fruitfully employed for (already planned) further investigations with higher frequency (monthly) youth unemployment data.

The use of an ANN in youth unemployment studies during a pandemic of the present scale is, to the best of the authors' knowledge, unprecedented. The study can be replicated with data from different countries and over longer horizons as new (real) data comes in.

## References

Agenda 2030 (2015). Transforming our world: the 2030 Agenda for Sustainable Development. Accessed at:

<https://sustainabledevelopment.un.org/post2015/transformingourworld>. Access date: 23.05.2020.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov and F. Csaki, eds., 2nd International Symposium on Information Theory (Akademia Kiado, Budapest).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Anders, U., & Korn, O. (1996). Model selection in neural networks. ZEW Discussion Papers, 96-21.

Anvik, C., & Gjelstad, K. (2010). Just Google it. *Forecasting Norwegian unemployment figures with web queries*, Center for Research in Economics and Management CREAM Publication, 11.



- Askatas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2), 107-120.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
- Bruno, G. S., Marelli, E., & Signorelli, M. (2014). The rise of NEET and youth unemployment in EU regions after the crisis. *Comparative Economic Studies*, 56(4), 592-615.
- Choudhry, M. T., Marelli, E., & Signorelli, M. (2012). Youth unemployment rate and impact of financial crises. *International Journal of Manpower*, 33(1), 76-95.
- Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, 1-5.
- Coibion, O., Gorodnichenko, Y., & Weber, M. (2020). *Labor markets during the covid-19 crisis: A preliminary view* (No. w27017). National Bureau of Economic Research.
- D'Amuri, F., & Marcucci, J. (2010). 'Google it!'Forecasting the US unemployment rate with a Google job search index.
- D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801-816.
- Derrick, T. R., & Thomas, J. M. (2004). Time series analysis: the cross-correlation function.
- De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496), 1513-1527.
- De Livera, A. M. (2010), Forecasting time series with complex seasonal patterns using 538 exponential smoothing, PhD thesis, Monash University.
- Demidova, O., & Signorelli, M. (2012). Determinants of youth unemployment in Russian

regions. *Post-Communist Economies*, 24(2), 191-217.

Di Stefano, E. (2019). Leaving your mamma: why so late in Italy?. *Review of Economics of the Household*, 17(1), 323-347.

Doerr, S., & Gambacorta, L. (2020). Identifying regions at risk with Google Trends: the impact of Covid-19 on US labour markets. Accessed at:

<https://www.bis.org/publ/bisbull08.pdf>. Accessed on: 26.05.2020.

EC Country Report (2020). European Commission Country Report Italy 2020. Accessed at [https://ec.europa.eu/info/sites/info/files/2020-european\\_semester\\_country-report-italy\\_en.pdf](https://ec.europa.eu/info/sites/info/files/2020-european_semester_country-report-italy_en.pdf).

Access date: 23.05.2020.

Eurostat (2020). Eurostat NEET quarterly data. Accessed at

[https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfsi\\_neet\\_q&lang=en](https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfsi_neet_q&lang=en). Access date:

23.05.2020.

Fellini, F. (1960). *La dolce vita* (No. Nd429). DVD VIDEO.

Fine, T. L. (2006). Feedforward neural network methodology. Springer Science & Business Media.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014.

Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment?. *Economic Modelling*, 30, 117-125.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297-307.

ILOSTAT (2020). Homepage. Accessed at: <https://ilostat.ilo.org/>. Access date: 23.05.2020.

- ILO Monitor (2020). ILO Monitor Third Edition, 29 April 2020. Accessed at: [https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/briefingnote/wcms\\_743146.pdf](https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/briefingnote/wcms_743146.pdf). Access date: 23.05.2020.
- Murata, N., Yoshizawa, S., & Amari, S. I. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE transactions on neural networks*, 5(6), 865-872.
- Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, 130, 114-122.
- Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the US unemployment rate using Google Trends. *Finance Research Letters*, 30, 103-109.
- Quintano, C., Mazzocchi, P., & Rocca, A. (2018). The determinants of Italian NEETs and the effects of the economic crisis. *Genus*, 74(1), 5.
- SDG Progress Report (2019). The Sustainable Development Goals Progress Report 2019. Accessed at: <https://unstats.un.org/sdgs/report/2019/>. Access date: 23.05.2020.
- Statista (2020). Search engines ranked by market share in Italy 2019. Accessed at: <https://www.statista.com/statistics/623043/search-engines-ranked-by-market-share-in-italy/>. Access date: 28.05.2020.
- UN (2020). UN Special Issue on COVID-19 and youth. Accessed at: <https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2020/04/YOUTH-FLASH-Special-issue-on-COVID-19-1.pdf> Access date: 23.05.2020.
- Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing?. *Technological Forecasting and Social Change*, 92, 132-139.
- Vinod, K., Varghese, N., Roy, S. B., & Syamaprasad, U. (2009). Significant enhancement of

the in-field critical current density of the MgB2 superconductor through codoping of nano-TiC with nano-SiC. *Superconductor Science and Technology*, 22(5), 055009.

Vinod, H. D. (2013). Maximum entropy bootstrap algorithm enhancements. *Available at SSRN 2285041*.

## **Appendix I**

ILOSTAT (2020) calculates the NEET rate as follows:

NEET rate (%) = (Youth – Youth in employment – Youth not in employment and education or training)/ Youth,

which can also be expressed as:

NEET rate (%) = ((Unemployed youth + Youth outside the labour force) – (Unemployed youth in education or training + Youth outside the labour force in education or training))/ Youth \* 100.