Journal of Data Intelligence, Vol. 2, No. 3 (2021) 348–367

SUSTAINABLE DEVELOPMENT GOAL RELATIONAL MODELLING AND PREDICTION: INTRODUCING THE SDG-CAP-EXT METHODOLOGY

YASSIR ALHARBI
Department of Computer Science, The University of Liverpool
Liverpool, L69 3BX, United Kingdom
Al-Mahd College, Taibah University
Medina 42353, Saudi Arabia
Yassir.Alharbi@liverpool.ac.uk

DANIEL ARRIBAS-BEL
School of Environmental Science, The University of Liverpool
Liverpool, L69 3BX, United Kingdom
darribas@liverpool.ac.uk

FRANS COENEN
Department of Computer Science,The University of Liverpool
Liverpool, L69 3BX, United Kingdom
coenen@liverpool.ac.uk

A methodology for UN Sustainable Development Goal (SDG) attainment prediction is presented, the Sustainable Development Goals Correlation Attainment Predictions Extended framework SDG-CAP-EXT. Unlike previous SDG attainment methodologies, SDG-CAP-EXT takes into account the potential for a causal relationship between SDG indicators both with respect to the geographic entity under consideration (intra-entity) and neighbouring geographic entities to the current entity (inter-entity). The challenge is in the discovery of such causal relationships. A ensemble approach is presented that combines the results of a number of alternative causality relationship identification mechanisms. The identified relationships are used to build multi-variate time series prediction models that feed into a bottom-up SDG prediction taxonomy, which is used to make SDG attainment predictions and rank countries using a proposed Attainment Likelihood Index that reflects the likelihood of goal attainment. The framework is fully described and evaluated. The evaluation demonstrates that the SDG-CAP-EXT framework can produce better predictions than alternative models that do not consider the potential for intra- and inter-causal relationships.

Keywords: Time Series Correlation and Causality, Hierarchical Classification, Time Series Prediction and Forecasting, United Nations Sustainable Development Goals

1. Introduction

Time series analysis is a significant task undertaken in many organisations in many different fields; example applications include budgeting, new product enrolment, outcomes of medical procedures, and many more [1, 2, 3]. The ability to make predictions for future values using historical values can help any organisation better plan for future events. Predicting the future value for a variable v using only historical data may sometimes produce precise forecasts provided v has no external influence. In such cases, an approach as simple as a rolling mean can be used [4]. However, in practice, the time series to be considered feature noise of various kinds [5], and/or maybe very short [6]. One example application domain where this is the case, and the focus of the work presented in this paper, is in the context of the United Nations (UN) Sustainable Development Goals (SDGs); a collection of 17 interlinked global goals, comprised of 169 different targets, designed to be a "blueprint for achieving a better and more sustainable future for all".[a] To monitor SDG attainment, the UN collects statistical data on a regular basis. This can be formed into a data set of some 200,742 short time series where each time series has a maximum of 22 values, from 2000 to 2021.

The challenges of SDG attainment prediction can be summarised as follows: (i) the time series to be utilised are short (maximum of 22 observations); (ii) the noisy nature of the data, which also features a lot of missing values, and which therefore needs an intensive amount of preprocessing and interpolation; (iii) the hierarchical nature of the data (geographical location $\rightarrow$ goal $\rightarrow$ target $\rightarrow$ indicator $\rightarrow \ldots$); (iv) the lack of specific attainment value (thresholds); (v) the computational complexity of causal inference in the context of the short SDG time series, and (vi) the absence of clear guidance on how to classify goal achievement based on progress.

In [7] the SDG time series data was used to make predictions regarding future SDG attainment under the assumption that each time series was independent. In [8] the SDG-CAP methodology was presented for predicting SDG attainment that took into consideration intra-geographic entity causal/correlation relationships. It was demonstrated that by using combinations of SDG times series, founded on observed correlations between time series pertaining to a given country, more accurate predictions could be made than in the case of [7]. In this paper, the SDG-CAP-EXT methodology is presented that takes into account the possibility of both intra- and inter-geographic entity relationships.

From the foregoing, the hypothesis considered in this paper is that it is possible to obtain a more accurate SDG attainment prediction by examining both the possible relationships between the SDG time series within a given country and with respect to neighbouring countries. As in the case of [7, 8] the idea is to conceptualise SDG attainment as a tree hierarchy where the nodes are SDG goals, targets, indicators, sub indicators and sub-sub-indicators. A predictor is then held at each leaf node, the results of which feed up the hierarchy to the root of the tree. Thus instead of building each leaf node predictor according to the relevant time series data (a one-to-one correspondence) as in [7], or as a set of correlated nodes within a single country as proposed in [8], it is proposed in this paper that it might be better if the time series data sets used to build the predictors were more comprehensive, in other words, founded on both intra- and inter-relationships. The challenge then is to determine these causal/correlation inter-relationships. In [8] time series were grouped by simply selecting the top five most closely related time series for each SDG leaf indicator. In this paper, a much more sophisticated approach is presented. A further disadvantage of the work presented in [7] and [8] was that the result was a Boolean, a given country either will or will not attain the UN SDG requirements. It is suggested here that, where a country is predicted not to attain the SDG requirements, it would be more helpful if a measure of how far away the country was from attainment. This paper thus also proposes the use of an Attainment Likelihood Index (ALI) produced using the forecasted results.

[a]UN resolution 71/313 adopted by the General Assembly on 6 July 2017

Given the above, this paper proposes the SDG Correlation/Causal Attainment Prediction Extended (SDG-CAP-EXT) methodology designed to address the disadvantages associated with the previous work presented in [7, 8], although the work in [7] and [8] provides an excellent forecasting benchmark. The main challenge is determining which time series are influenced by which other time series in a given region. This can be done by hand given a domain expert and sufficient time resource. However, automating the process is much more desirable. The work presented in this paper provides a potential solution to this problem.

In the context of the proposed SDG-CAP-EXT methodology, this paper makes three contributions:

1. An investigation into mechanisms whereby relationships between short time series can be identified in a more sophisticated manner than suggested in [8].
2. A mechanism whereby both an intra-geographic region and inter-geographic region causality/correlation relationships can be identified so as to identify a related sets of time series that can be used for prediction purposes in a multi-variate setting.
3. The usage of the ALI metric to rank countries based on their SDG attainment likelihood.

The rest of the paper is organised as follows. In the following section, Section 2, a brief literature review of the previous work underpinning the work presented in this paper is given. The SDG application domain and the SDG time series data set is described in Section 3. The proposed SDG-CAP-EXT methodology is then described in Section 4. The evaluation of the proposed methodology is presented in Section 5. The paper concludes with a summary of the main findings, and several proposed direction for future research, in Section 6.

2. Literature Review

As noted above, the challenges that the proposed SDG-CAP-EXT methodology seeks to address include: (i) short time series forecasting and (ii) time series causal inference. Previous work in these two areas is therefore considered in the first two sub-sections of this literature review. The literature review is completed with some discussion of previous work directed at SDG forecasting.

2.1. *Short time series forecasting*

Prediction using short time series is challenging because it is difficult to make a significant off-sample assessment, or cross-validation, given a small number of observations [6]. A range of methods has been proposed to address this issue [11]. However, many of the proposed solutions still require no less than 50 observations. With regard to SDG data, the sample size is below 20 observations. The FBProphet time series forecasting tool requires a minimum of 3 observations and was used in [7] for the purpose of SDG attainment prediction where it was demonstrated that FBProphet produced a better prediction accuracy than two alternatives, Auto-Regressive Moving Average (ARMA) [13] and Auto-Regressive Integrated Moving Average (ARIMA) [11].

However, FBProhpet is a uni-variate predictor; given that this paper's focus is prediction using sets of causal-related time series, a multi-variate approach is required; in [7] it was assumed that each SDG time series was independent of all other time series, hence univariate prediction. A multi-variate time series forecasting model, using Long Short Term Memory (LSTM) networks, was presented in [12]. The LSTM model demonstrated better overall performance than the ARMA and ARIMA alternatives [13, 11]. The LSTM model was adopted in [8] for multi-variate SDG attainment forecasting with respect to intra-country relationships. More generally, LSTM models have been widely adopted with respect to many real-life applications such as weather forecasting [1] and stock market [3] prediction. With respect to the work presented in this paper, an Encoder-Decoder LSTM was used [14]. LSTM typically perform better when large data sets are used, but also seems to perform well when a large number of short time series are used in a multi-variate setting.

2.2. *Time series causal inference*

Causal inference is concerned with the process of establishing a connection (or the lack of a connection) between events or instances. Given two candidate time series, $A = \{a_1, a_s, \dots, a_n\}$ and $B = \{b_1, b_2, \dots, b_m\}$, where we wish to establish that B is causality-related to A, this is typically established using a prediction mechanism that uses the "lag" $\{b_1, \dots, b_{m-1}\}$ to predict a_n. We then compare the predicted value for a_n with the known value, for example using the Root Mean Square Error (RMSE). If the two values are close then this indicates that "time series A is causality-related to time series B".

A number of mechanisms can be adopted to identify causal inference. In [8] four mechanisms were considered: (i) Granger Causality [15, 16, 17], (ii) the Temporal Causal Discovery Framework (TCDF) [20], (iii) Pearson Correlation [21] and (iv) the Least Absolute Shrinkage Selector Operator (LASSO) [22]. TCDF did not work well, because it was found to require longer time series than those available within the SDG data. With respect to the work presented in this paper Granger Causality, Pearson Correlation and LASSO were also considered together with two further mechanisms: the Mann-Whitney U Test and Dynamic Time Warping (DTW). Each of the five causality mechanisms considered with respect to the proposed SDG-CAP-EXT methodology are discussed in some further detail in sub-sections 2.2.1 to 2.2.5 below.

For the evaluation presented later in this paper the operation of the proposed SDG-CAP-EXT methodology was compared with the SDG-CAP methodology described in [8] and the option of using Principle Component Analysis (PCA). SDG-CAP has already been discussed, PCA is briefly described in Sub-section 2.2.6 below.

2.2.1. *Granger Causality*

Granger Causality (GC) is one of the most widely used causal inference mechanisms found in the literature [15, 16, 17]. It was introduced in the 1960s and is calculated as shown in Equation 1 where: (i) A and B are time series, (ii) a and b are the laggs of A and B, (iii) t is the current time step and (iv) e is a residual error. The idea is that if time series A "granger causes" time series B, then the past values of A should contain helpful information to forecast B in a manner that would be better than when forecasting B without the inputs from A. The variation of GC that was used with respect to the research presented in this paper is the Statsmodels variation [18]. GC has been used previously in the context of SDG prediction, for example, in [19] and, as noted above, in [8]. In [19] 20,000 pairs of SDG time series that featured causal relationship were found.

$$A_t = a_1 A_{t-1} + b_1 B_{t-1} + e \tag{1}$$

2.2.2. *Pearson Correlation*

Pearson Correlation [21] has been used to measure the correlations between any given pair of time series. The mechanism assumes linearity of the data. This assumption holds with respect to many SDG time series that are typically linearly spaced.

2.2.3. *Lasso*

Lasso [22] is an L1 regularisation technique frequently used to reduce high dimensionality data, which can also be employed to establish the existence of causality between variables [23, 24, 25, 26]. LASSO reduces the dimensionality of the input data set by penalising variances to zero, thus allowing irrelevant variables to be removed. Equation 2 shows the LASSO cost function. Inspection of the equation indicates that the first part is the $squarederror$ function, whilst the second part is a penalty applied to the regression slope. If λ is equal to 0, then the function becomes a normal regression. However, if λ is not 0, coefficients are penalised accordingly, leaving only coefficients that can explain the variance in the data.

$$Lasso\ Cost\ Function = \sum_{i=1}^{n} \left(y_i - \sum_j x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

2.2.4. *Mann-Whitney U Test*

The Mann-Whitney U Test [27] is the fourth causal inference mechanism used in this paper. The test is used to determine if any two time series are statistically different. It is a non-parametric test (unlike, for example, LASSO).

2.2.5. *Dynamic Time Wrapping DTW*

The DTW measure [9] is a time series similarity measure; which, unlike Euclidean distance [28], does not require the time sequences to be of the same length. Thus it can generate a similarity measurement between two sequences of various lengths. The test is used to determine if any two pairs of time series are similar and therefore can be said to be correlated.

2.2.6. *Principle Component Analysis*

PCA [10] is a high dimensionality reduction and feature extraction technique. PCA reduces the number of features in a data set whilst retainig all significant information. The significance with respect to the work presented in this paper is that it was considered as an alternative method for selecting time series to be used for multi-variate forecasting with respect SDG attainment.

2.3. ***Sustainable Development Goals Forecasting***

Previous work directed at forecasting SDG attainment can be divided into two main categories: (i) single target forecasting or (ii) multiple target forecasting. The first is directed at forecasting with respect to an individual SDG or specific geographical location. Much existing work falls into this category. Examples can be found in [29] and [30] where forecasting was directed at a specific region (Ukraine) or a specific SDG (electricity supply), respectively. An example of the second category can be found in the context of The International Future Scenarios framework[b] [31], an approach for forecasting SDG variables in a multi-variate manner. However, it uses predefined scenarios based on several methods, such as a regression model in the case of health goals [31].

This approach limits the input of the independent variables to predefined scenarios. To the best knowledge of the authors the proposed SDG-CAP-EXT methodology is the only framework/methodology where by both intra- and inter-geographic entity relations can be used for holistic SDG attainment prediction.

3. The United Nation's Sustainable Development Goal Agenda

At the beginning of the 20Th century, the UN announced its first development plan for a set of eight development goals, listed in Table 1, to be attained on or before 2015 for all member states [32]. These were referred to as the Millennium Development Goals (MDGs). In 2015 the UN built on the success of the MDG objectives and introduced the Sustainable Development Goals (SDGs), listed in Table 2, to be achieved by 2030 [33]. Each SDG has several targets (sub-goals) and indicators (sub-sub-goals) associated with it, each linked to an attainment threshold of some kind (to be achieved by a given date). For example, for SDG 1, "No Poverty", which comprises six sub-goals, the extreme poverty threshold is defined as living on less than 1.25 USD a day. In this paper we indicate SDG sub-goals using the notation $g.t.i.\ \ldots$, where g is the goal number, t is the target number, i is the indicator number, and so on. For example, SDG 2.22 indicates target 22 of SDG 2. The UN has made available the MDG/SDG data collated so far .[c]

1.	To eradicate extreme poverty and hunger.
2.	To achieve universal primary education.
3.	To promote gender equality and empower women;
4.	To reduce child mortality.
5.	To improve maternal health.
6.	To combat HIV/AIDS, malaria, and other diseases.
7.	To ensure environmental sustainability.
8.	To develop a global partnership for development.

Table 1. The eight 2000 Millennium Development Goals (MDGs)

In [7] the complete set of SDGs and associated target and indicators was conceptualised as a taxonomic hierarchy, as shown in Figure 1. In the figure, the root node represents the complete set of SDGs for a country (geographic region), the next level the seventeen individual SDGs, then the sub-goals referred to as "targets", the sub-sub-gaols referred to as "indicators"

[b] https://pardee.du.edu/
[c] https://unstats.un.org/SDGs/indicators/database/

1.	No Poverty.
2.	Zero Hunger.
3.	Good Health and Well-being.
4.	Quality Education.
5.	Gender Equality.
6.	Clean Water and Sanitation.
7.	Affordable and Clean Energy.
8.	Decent Work and Economic Growth.
9.	Industry, Innovation and Infrastructure.
10.	Reduced Inequality.
11.	Sustainable Cities and Communities.
12.	Responsible Consumption and Production.
13.	Climate Action.
14.	Life Below Water.
15.	Life on Land.
16.	Peace and Justice Strong Institutions.
17.	Partnerships to Achieve the Goal.

Table 2. The seventeen 2005 Sustainable Development Goals (SDGs)

and so on. This taxonomy was also used on [8] and is also adopted with respect to the work presented in this paper.

The UN SDG data set comprises a single (very large) table with the columns representing a range of numerical and categorical attributes and the rows representing single observations related to individual SDG sub-goals sub-sub-goals. Each row is date stamped. The data set features 283 different geographical regions. For each region there were, as of February 2021, up to 2,446 different time series The maximum length of a time series was 22 points, covering 22 year's of observations, although a time series featuring a full 22 observations was unusual; there were many missing values. In some cases data from earlier years were also included. In the context of the research presented in this paper, only data from the year 2000 to 2017 was considered; 165,227 time series in total. An example is given in Figure 2 where two time series are given describing the proportion of the population of Algeria, rural and urban, with access to electricity; note that correlations can be observed between the two. By applying time series analysis to the data, trends can be identified for prediction/forecasting purposes.

Further challenges associated with SDG data are missing values and the range of measurement units used. The number of missing values in the SDG data set presented a particular problem. The total theoretical number of observations (time series points) in the data was 4,416,324, while the actual number was 1,573,099; in other words, the data featured 2,843,225 missing values (64.38%). This is illustrated in Figure 3 where the percentage of values recorded compared to the expected total number of values is presented with respect to the North Africa geographic region (note the large number of missing values from 2017 onwards). In some case, the missing data could be explained, for example, because observations were only made following a five-year cycle. However, in most cases, the missing values seemed to be missing in what could only be described as a random manner. With respect to the units used, the SDG data provides measurements of life events in a variety of domains. Thus, there are 44

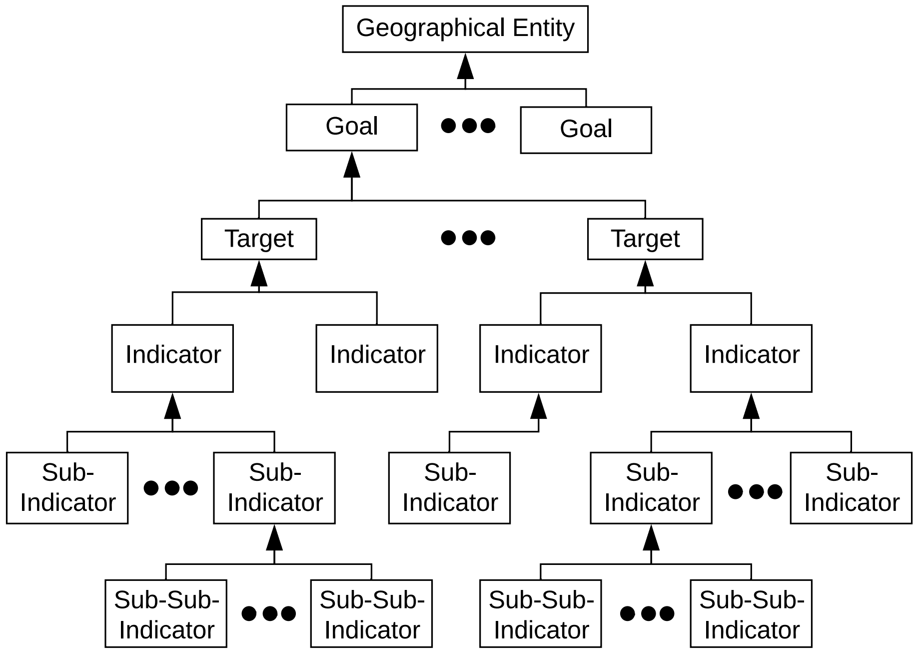

Fig. 1. The SDGs taxonomy proposed in [7]

different measurement units used. Figure 4 presents a plot of the number of values per unit of measurement within the SDG data. From the figure, it can be seen that the majority of the data uses percentage as the unit.

Given the foregoing, the SDG data required considerable preprocessing. Figure 5 presents a schematic of the preprocessing framework that was adopted so that the SDG data was in the time series format required for application of the proposed SDG-CAP-EXT methodology. It should be noted here that the data preprocessing, although extensive, only needs to be done once, or at last only once for each update of the SDG data. From the figure, the preprocessing is conducted in five steps: (i) transposing, (ii) taxonomy generation, (iii) filtering, (iv) scaling and (v) imputing. The preprocessing commences with the transposing of the raw 21×43 row-column format data for each time series associated with a leaf node in the taxonomy given in Figure 1 to a 1×24 row-column format, where each row is of the form:

$$\langle \mathrm{GR}, \mathrm{G}, \mathrm{T}, \mathrm{I}, \mathrm{D}, t_0, \ldots, t_{21} \rangle \tag{3}$$

where GR is a Geographical region, G is an SDG, T is a target, I is an indicator, D is the descriptor for the time series, and t_0 to t_{21} are the associated time series values for the years 2000 to 2021. By aggregating the data in this manner, no data is lost, and the tree shape is maintained. The indicator IDs provide a unique key to the level three nodes of the tree (see Figure 1), while the descriptors provide a unique reference to the time series at the leaf nodes (sub-indicators or sub-sub-indicators). An example record is given in Table 3. The descriptor SH_DYN_MORTN (“ Under-five deaths (number)”) provides a unique reference to the time series, which, with a little practice, can be readily interpreted by a human reader.

The data is then filtered based on the number of missing values. Any time series with more

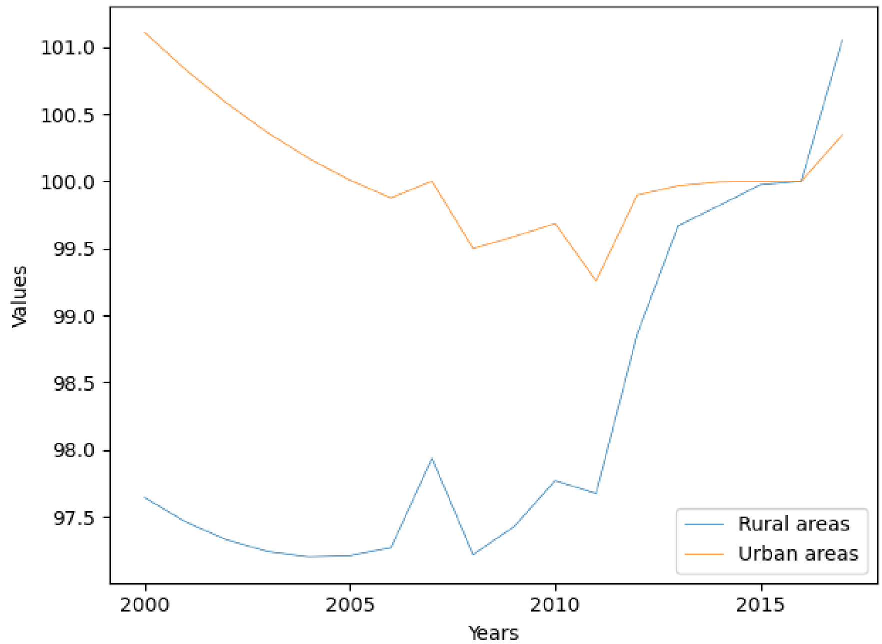

Fig. 2. Proportion of the population of Algeria with access to electricity (2000-2017)

GR	**G**	**T**	I	**Unique Code**	2000	2001	....	2019	2020	2021
Algeria	3	2	3.2.2	SH_DYN_MORTN_BOTHSEX_5Y	24415	23622	....	23598	NA	NA

Table 3. Example of a transposed data record

than 15 missing values or featuring irregularities, such as the presence of five zeros in a row, is deemed to be noisy data and is put to one side in a set $T_{noise} = \{T_1, T_2, \dots\}$ (the "Noisy Data" drum in Figure 5). The rest of the data, T, will then be scaled using RobustScaler [34] so that all data uses a single unified scale. The next step in the preprocessing was to impute missing values. Experiments were conducted using four different imputation methods: (i) Linear, (ii) Krogh, (iii) Spline and (iv) Pchip [35, 36, 37]. The aim was to identify the most appropriate imputation method. The experiments were conducted using complete time series only. Then s-random deletion of the data was applied to simulate the missing values situation, but with a ground truth in that the missing values were known. The four different candidate imputations algorithms were then applied. Root Mean Square Error (RMSE) measurement was used to ascertain the performance of the different imputation methods. The results are presented in Table 4 with respect to the time series for Target 3.2 for Egypt. From the table, it can be seen that the worst average performance was associated with the Krogh method, while the Spline method produced the best average performance. Thus the Spline method was chosen to be incorporated into the proposed SDG-CAP-EXT methodology. In practice we found it appropriate to use data from 2000 to 2017 inclusive because of the large number of missing values from 2018 onward, as illustrated previously in Figure 3. The final output was a set $T_{clean} = \{T_1, T_2, \dots\}$.

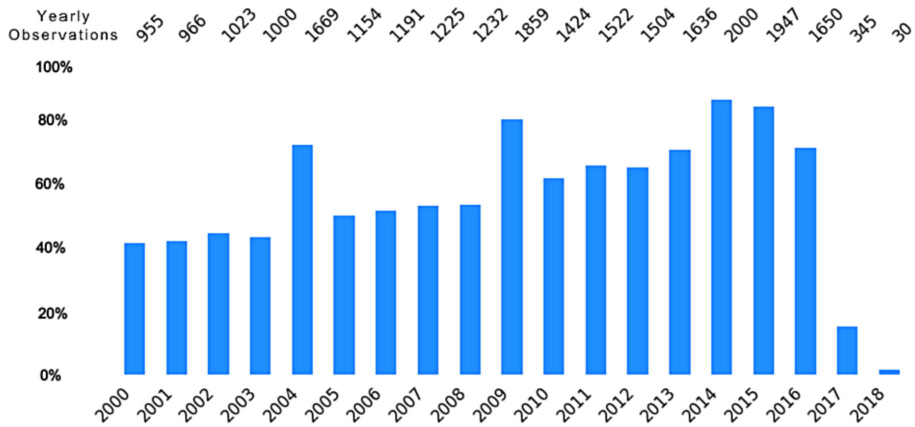

Fig. 3. Number of observations in the SDG data for the North Africa region per year

4. The Sustainable Development Goal Correlation/Causal Attainment Prediction Extended (SDG-CAP-EXT) Methodology

This section presents the SDG-CAP-EXT methodology. The methodology's objective is to predict whether a country x will reach its SDG attainment targets. A schematic of the methodology is given in Figure 6. The input is the SDG time series data D for a geographic region; for example, North Africa or Central America. The output is a set of predictions for the countries within the given region. From the figure, it can be seen that the methodology features five principal stages: (i) Data preprocessing (Stage 1), (ii) Causation identification (Stage 2), (iii) Prediction (Stage 3), (iv) Taxonomy population (Stage 4) and (iv) Visualisation (Stage 5). Recall that each time series for a given country is associated with a leaf node in the SDG taxonomy and that we wish to generate predictors with respect to each leaf node to make predictions regarding attainment of the indicator represented by the leaf node that can then be used to populate the SDG taxonomy tree, for each country in the given region, to make overall predictions regarding SDG attainment for each country.

The data preprocessing stage was discussed in the previous section where it was noted that, as a result of preprocessing, the SDG input data D is partitioned into two sets based on the number of missing values in each time series, T_{noise} where we have 3 points or less, and T_{clean} otherwise. The distinction was that the first could be used for causality/correlation relation identification, while the second could not. However, as indicated in the schematic given in Figure 6, we do not entirely discard T_{noise}, this collection is processed further and partitioned into T_{noise_3} and $T_{noise_{\leq 2}}$, where T_{noise_3} comprises time series with 3 points and $T_{noise_{\leq 2}}$ comprises time series with 2 points or less. The significance is that time series with three points could still be used to make predictions, although it could not be used for the purpose of discovering causality/correlation relationships.

In Stage 2, causation identification, every time series $T_t \in T_{clean}$ was compared to every other $T_s \in T_{clean}$ using a number of causation/correlation mechanisms. With respect to the evaluation presented later in this paper (Section 5) five causality relationship identification mechanisms were used; as indicated in Figure 6. The five mechanisms were: Granger Causality, LASSO regression, Pearson's Correlation, DTW and the Mann-Whitney U-Test. Both

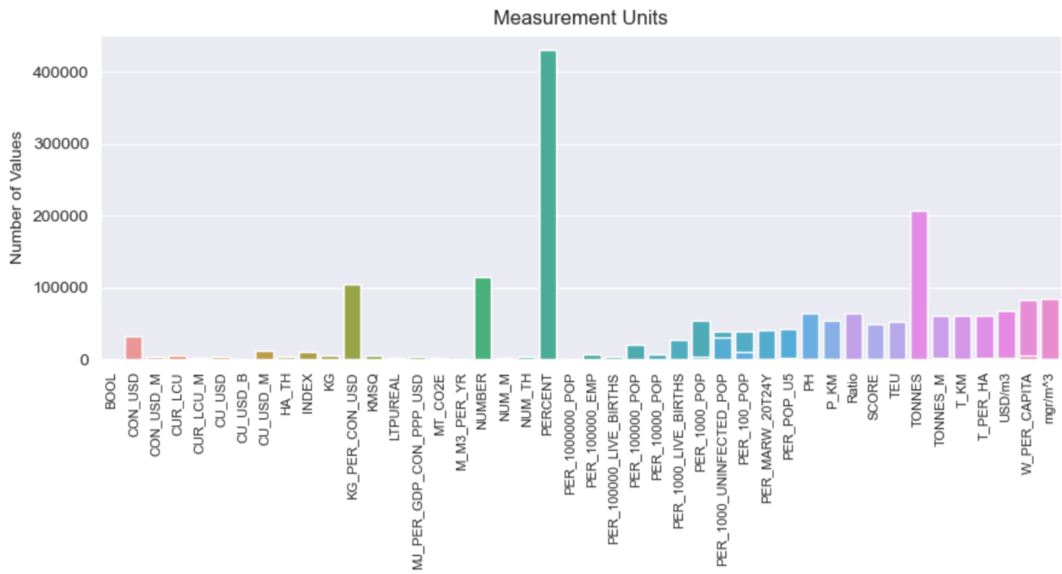

Fig. 4. Number of observations per measurement unit

Granger causality and the U-Test are hypothesis tests and produce a p-value between 0 and 1; Pearson's, DTW, and LASSO produce R^2 values. The intuition behind using five different causality/correlation relationship identification mechanisms was that each would operate differently. Hence, it was hypothesised, a combination of a number of different relationship identification mechanisms would work better than when those mechanisms were used in isolation. Note that the methodology will work equally well with four or six mechanisms. For future work, the authors intend to conduct further experiments using both less than, and more than, five mechanisms. The outcome for each mechanism is a $n \times n$ *causality matrix*, where n is the number of time series in T_{clean}. The cells in the matrix will hold causality values. Note that we conduct normalisation so that all values are within the range $[0.00, 1.00]$. The matrix can be visualised in the form of a heat map. A fragment of such a heat map, generated using LASSO and the geographic area Egypt, is given in Figure 7. The darker the colouring of each cell, the greater the LASSO R^2 weighting and the greater the causality/correlation. The leading diagonal is where time series are compared with themselves; hence we expect this to have the highest score.

The next process in Stage 2 is to use the causality matrix to identify the groups of time series that can be used for multi-variate time series SDG attainment prediction with respect to individual indicators. For each target time series, T_t, associated with an SDG indicator (represented by a column or row in the causality matrix), for which we wish to build a predictor the causality/correlation values are collated into four bins, $[0.00, < 0.25]$, $[0.25, < 0.50]$, $[0.50, < 0.75]$ and $[0.75, 1.00]$. We then rank the values in the $[0.75, 1.00]$ bin and select the top ten provided there are 10 or more values; it is possible that bin $[0.75, 1.00]$ is empty, but we have not found this to be the case with respect to any of our experiments. Thus, given that we advocate using five causality/correlation mechanisms, we can have five groups of ten time series, fifty in total. This is too many for our proposed multi-variate forecasting. Thus

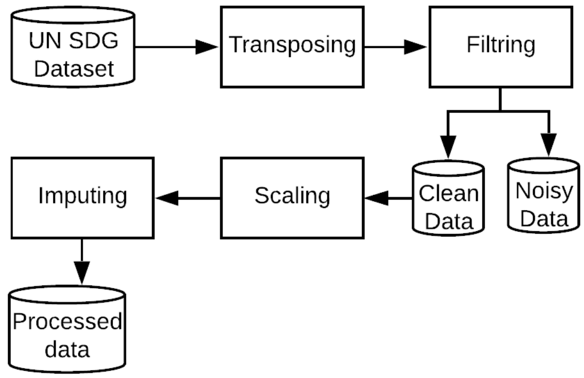

Fig. 5. Schematic of the adopted SDG data preprocessing framework

a voting mechanism is applied. Each time series is allocated a number of votes according to how many groups it appears in. Thus each time series will have a maximum of five votes and a minimum of one vote. We choose the top ten time series with the most votes out of the time series with two votes or more; thus, only time series that appeared in more than one time series group will be considered. It may be the case that there are less than ten time series that can be selected, in which case we select as many as we can. To distinguish between time series with equal numbers of votes, we refer back to the relevant causer matrices and determine average values which can be compared. The selected time series were placed in a "Causer Table", T_{causer}, where the rows represent the indicators represented by the leaf nodes in the taxonomy, each of which in turn will be represented by a target time series T_t; and the columns the maximum of ten time series that are related to T_t.

The next stage in the workflow shown in Figure 6 is the multi-variate prediction model generation stage, Stage 3. For each $T_t \in T_{causer}$, a hybrid-multi-variate time series forecasting model was built. The model to be associated with the leaf node that T_t represents. A range of tools and techniques are available whereby such a model can be constructed. However, for the evaluation presented later in this paper, a multi-variate LSTM-Encoder-Decoder (Enc-Dec) [38] was used.

Recall that some of the leaf nodes will be associated with the time series held in T_{noise_3} and $T_{noise_{\leq 2}}$. It is still possible to make predictions using a three point times series, although less than ideal. For the proposed SDG-CAP-EXT methodology, it is advocated that FBProphet is used with respect to the time series in T_{noise_3}. With respect to the time series in $T_{noise_{\leq 2}}$ nothing can be done, and the leaf nodes are marked with "unkonwn" (no prediction). From the evaluation reported on in the following section it was found that using the proposed SDG-CAP-EXT methodology prediction models for the majority leaf nodes were generated (but not all).

The next stage in the methodology, Stage 4, is to populate the taxonomy. For this purpose, a similar process to that described in [7] was adopted. Where possible, using either a generated LSTM model or FBProphet, predictions were made for the desired year y (provided as input to the methodology) and all the intervening years from the last year included in the input set D. Recall that every target has a threshold and a deadline for when that threshold should be

SDG 3.2	Algorithm			
	Linear	Krogh	Spline	Pchip
1	0.212	3938.998	0.536	**0.168**
2	3.952	14421.321	1.959	**1.864**
3	0.047	**0.018**	0.047	0.031
4	**0.000**	102.356	4.089	**0.000**
5	0.251	37.856	0.250	**0.054**
6	**0.861**	2687.759	1.330	1.125
7	0.559	9.548	**0.374**	0.731
8	0.042	0.727	0.042	**0.017**
9	1.820	70.773	2.596	**1.250**
10	6.924	1005.504	**1.456**	28.018
11	5.707	14196.115	**2.320**	20.260
12	0.036	0.025	0.095	**0.020**
13	0.032	0.175	0.032	**0.015**
14	**0.256**	13.235	**0.256**	0.299
15	0.063	0.148	0.064	**0.017**
16	2.167	21.497	2.167	**1.175**
Ave. Error	1.433	2281.629	**1.101**	3.440
Stand. Dev.	2.198	4830.316	**1.223**	8.223

Table 4. RMSE comparison of imputation methods used to generate missing values in the SDG data (best results in bold font)

achieved. For the work presented here, the threshold values used in [7] were also used. The predicted values for each indicator were used to classify the indicator using three classes: (i) T (True) for indicator met on or before the deadline; (ii) F (False) for indicator not met on or before the deadline; and (iii) *unknown* to indicate that the time series in question belongs to $T_{noise_{\leq 2}}$.[d]

The ALI values were then calculated with respect to each target time series T_s held in T_{clean} using Equation 4, where $|class = T|$ is the number of target time series classified as true and $|class = F|$ is the number of target time series classified as false. Thus an ALI value of 1 will indicate that all the leaf node indicators in the taxonomy have been met. The ALI value is particularly useful where a country has failed to meet its SDGs as it indicates how far the country is away from meeting its SDGs. Note that the ALI concept can also be used to indicate how far a country is away from meeting a particular SDG, Target or Indicator.

The final stage in the methodology, Stage 5, was to visualise the outcome in an easy to interpret manner. This was achieved by providing: (i) a "country table", (ii) a visualisation of the populated taxonomy, and (iii) a world map with the ALI values plotted using a colour coding. A country table is a detailed view of a country's status with respect to the prediction's result. Examples of each are given in Section 5 below.

$$ALI = \frac{|class = T|}{|class = T| + |class = F|} \tag{4}$$

[d]This class label was also used in other circumstances, for example where it was unclear whether a threshold had been met or not because of a poorly defined threshold.

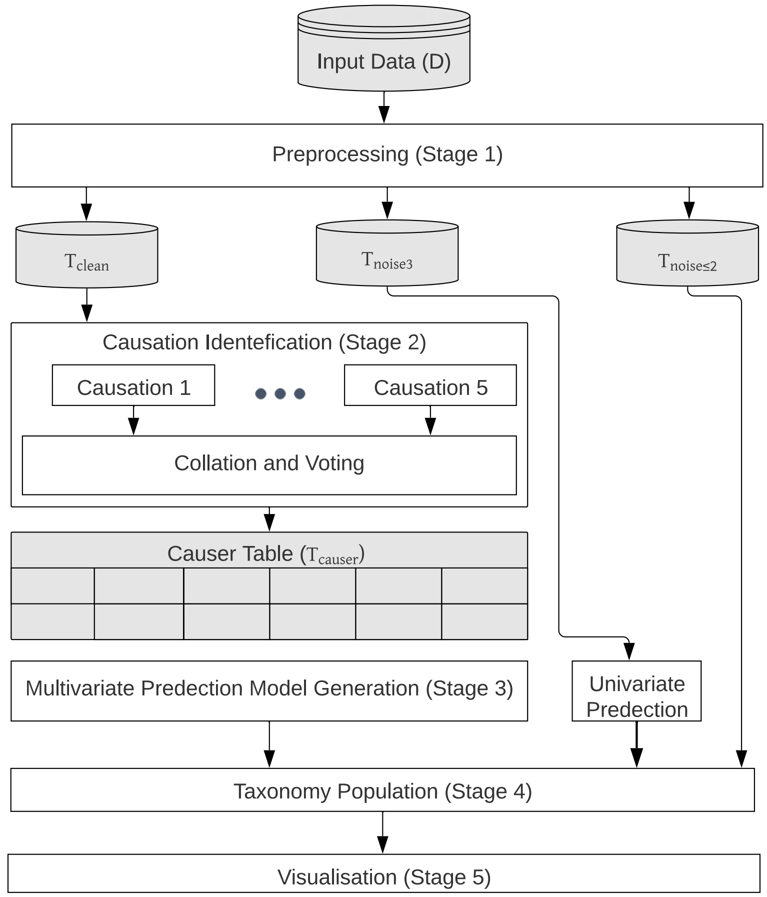

Fig. 6. Schematic of the SDG Causality/Correlation Attainment Predict Extended (SDG-CAP-EXT) Methodology

5. Evaluation

The proposed SDG-CAP-EXT methodology was evaluated by comparing its operation with the SDG-CAP methodology presented in [8] and the univariate approach presented in [7]. Recall that the SDG-CAP methodology presented in [8] was directed at identifying correlations within individual regions, while the approach presented in [7] assumed every time series was independent (no causality/correlation between time series). In addition, the proposed SDG-CAP-EXT methodology was compared with the operation of the methodology using only a single causality/correlation identification mechanisms, in other words using Granger, Pearson, Lasso, Mann-Whitney and DTW in isolation; and the operation of the proposed methodology if PCA was applied to select the "best" time series from the identified time series. As noted earlier, an inspection of the time series data indicated that in many cases there was no data for the years 2018 onward; hence eighteen point time series were used covering the years 2000 to 2017

For the evaluation, two sets of geographical entities were used. North Africa (Egypt,

Time Series	SDG 1_12	SDG 1_13	SDG 1_15	SDG 1_16	SDG 1_17	SDG 1_20	SDG 1_22	SDG 1_25	SDG 1_26	SDG 1_27	SDG 1_28	SDG 1_30	SDG 1_31	SDG 1_32	SDG 1_34
SDG 1_12	1.000	0.410	0.328	0.154	0.267	0.372	0.640	0.877	0.746	0.738	0.744	0.745	0.744	0.746	0.877
SDG 1_13	0.410	1.000	0.354	0.598	0.395	0.132	0.509	0.400	0.520	0.520	0.520	0.520	0.520	0.520	0.400
SDG 1_15	0.328	0.354	1.000	0.000	0.037	0.021	0.167	0.448	0.228	0.222	0.227	0.227	0.226	0.228	0.444
SDG 1_16	0.154	0.598	0.000	1.000	0.649	0.288	0.493	0.116	0.428	0.436	0.430	0.429	0.430	0.428	0.117
SDG 1_17	0.267	0.395	0.037	0.649	1.000	0.165	0.739	0.218	0.651	0.663	0.654	0.653	0.656	0.652	0.219
SDG 1_20	0.372	0.132	0.021	0.288	0.165	1.000	0.353	0.290	0.372	0.371	0.372	0.372	0.372	0.372	0.294
SDG 1_22	0.640	0.509	0.167	0.493	0.739	0.353	1.000	0.602	0.890	0.894	0.891	0.891	0.892	0.891	0.603
SDG 1_25	0.877	0.400	0.448	0.116	0.218	0.290	0.602	1.000	0.722	0.712	0.720	0.720	0.719	0.722	0.976
SDG 1_26	0.746	0.520	0.228	0.428	0.651	0.372	0.890	0.722	1.000	0.928	0.928	0.928	0.928	0.928	0.723
SDG 1_27	0.738	0.520	0.222	0.436	0.663	0.371	0.894	0.712	0.928	1.000	0.928	0.928	0.929	0.928	0.713
SDG 1_28	0.744	0.520	0.227	0.430	0.654	0.372	0.891	0.720	0.928	0.928	1.000	0.928	0.928	0.928	0.721
SDG 1_30	0.745	0.520	0.227	0.429	0.653	0.372	0.891	0.720	0.928	0.928	0.928	1.000	0.928	0.928	0.722
SDG 1_31	0.744	0.520	0.226	0.430	0.656	0.372	0.892	0.719	0.928	0.929	0.928	0.928	1.000	0.928	0.720
SDG 1_32	0.746	0.520	0.228	0.428	0.652	0.372	0.891	0.722	0.928	0.928	0.928	0.928	0.928	1.000	0.723
SDG 1_34	0.877	0.400	0.444	0.117	0.219	0.294	0.603	0.976	0.723	0.713	0.721	0.722	0.720	0.723	1.000

Fig. 7. A fragment of a heat map produced from the causality metrics generated using LASSO analysis and the geographic area Egypt [8]

Tunisia, Libya, Morocco, and Sudan) with a $|T_s| = 2300$, and Central America (Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, and Panama) $|T_s| = 5136$ (thus 7436 time series in total). The data sets were divided into a training and test sub-sets. The training set comprises the first 15 points, and the test set the last three.

The same experimental setup was used with respect to both data sets. The experiments were conducted using a windows machine running Python 3 on a Ryzen 9 processor with 40 Gig of ram and Nvidia RTX2060 GPU. For each time series, predictions were made using the different mechanisms itemised above, and a Root Mean Square Error (RMSE) value obtained. The RMSE value was calculated by comparing the predicted 2015, 2016 and 2017 values with the known values and an RMSE value derived. The total run time for the experiment was 266.64 hours (indicating the complexity of the evaluation). Some example results are presented in Table 5. In the table, the individual leaf node target time series are listed in column one using a numerical sequential numbering. The average RMSE values and associated Standard Deviation (SD) values are given in the last two rows. From the table, it can be seen that the SDG-CAP-EXT methodology produced the best result. Overall, it can also be observed that the proposed SDG-CAP-EXT methodology is well able to handle short time series.

	Regional Segments							**Single Country**	
Time Series #	**Lasso**	**DTW**	**Granger Causality**	**Pearson's Correlations**	**SDG-CAP-EXT**	**U_test**	**PCA**	**Univariate**	**SDG-CAP**
1	0.20	0.82	0.15	0.22	**0.12**	0.17	0.68	1.31	0.80
2	**0.15**	0.78	0.20	0.22	0.18	0.18	0.73	1.70	1.05
3	0.48	1.09	**0.35**	0.67	0.44	0.65	0.77	1.95	1.08
4	**0.15**	0.69	0.18	0.18	0.24	0.21	2.50	0.24	1.03
5	0.40	1.32	0.23	**0.22**	**0.22**	0.25	1.85	0.97	1.05
6	1.05	2.05	**0.62**	0.74	**0.62**	0.71	1.59	2.32	1.08
7	0.38	1.27	0.15	**0.07**	0.10	0.18	1.16	1.25	1.04
8	0.51	1.67	0.73	**0.49**	0.77	0.76	1.03	2.33	1.01
...	...	...	...	...	...	...	...	...	...
7433	0.68	0.70	**0.65**	2.00	0.67	0.70	1.69	2.20	0.96
7434	0.27	0.24	**0.21**	0.24	**0.21**	0.24	**0.21**	0.95	0.91
7435	2.00	**0.00**	0.03	0.01	**0.00**	**0.00**	2.00	1.84	1.04
7436	0.37	0.39	2.00	**0.31**	0.38	0.39	0.82	2.40	0.99
Averages	1.03	1.23	0.99	1.01	**0.85**	1.01	1.73	1.73	0.93
STD	1.68	1.52	1.57	1.39	**1.65**	1.66	1.04	2.66	1.84

Table 5. Example RMSE results using the proposed SDG-CAP-EXT methodology and a range of comparitor methodologies, (best results in bold font)

Using the SDG-CAP-EXT methodology the SDG taxonomy (Figure 1) will be populated for each country and the predicted values at the leaf nodes used as input to the parent nodes to make a final prediction as to whether the country in question will meet its SDGs. The results with respect to Indicator 3.2.1, "Under five child mortality", are presented in Table 6. In the table, the "Current" column gives the 2015 percentage of under five child deaths over the number of children under five, the "threshold" column gives the threshold required for this indicator to be met, less than or equal to 12%, the "prediction" column gives the predicted percentage (generated using the SDG-CAP-EXT methodology); and the "Result" column whether the indicator has been met or not. Inspection of the table shows that all countries considered in the evaluation, except Nicaragua, will meet indicator 3.2.1 on or before the deadline of 2030.

Geo Area	Country	Current	Threshold	Prediction	Result
Central America	Belize	4.7	<=12	4.86	Met
	Honduras	0.6	<=12	0.62	Met
	Costa Rica	0.5	<=12	0.51	Met
	El Salvador	0.4	<=12	0.36	Met
	Guatemala	1.4	<=12	1.94	Met
	Mexico	0.4	<=12	0.49	Met
	Nicaragua	19.3	<=12	19.87	Not Met
	Panama	0.5	<=12	0.77	Met
North Africa	Algeria	21.4	<=12	0	Met
	Egypt	20.8	<=12	0	Met
	Libya	11.4	<=12	5.88	Met
	Morocco	21.9	<=12	0	Met
	Sudan LDC	46.5	<=12	0	Met
	Tunisia	14.8	<=12	0	Met

Table 6. SDG attainment results with respect to Indicator 3.2.1, "By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births"

A fragment of the Country Table for Belize, produced using the proposed SDG-CAP-EXT methodology is given in Table 7. Each row represents a target time series. The first column gives the indicator ID made up of the G, T and I values with respect to the adopted time series format ($\langle \dots, G, T, I, \dots \rangle$) presented in Section 3. The second column gives the text description for the indicator taken from the raw SDG data. The "Descriptor" column gives the sub-indicator description for the leaf node target time series, the variable D with respect to the adopted time series format given in Section 3. The "Initial Value" column gives the current time series value (2015 in the case of the evaluation presented here). The "Predicted" column, as the name suggests, gives the predicted value for the target time series. The "Result" column gives the classification produced using the SDG-CAP-EXT methodology. Recall that this can be: (i) T, the sub-indicator (represent by the target time series) has been met, or is predicted to be met, on or before the deadline; (ii) F, the sub-indicator is predicted not to be met by the deadline; or (iii) *unknown*, either because the threshold is not clearly defined by the UN or the length of the associated target time series is less than 2 (the set

$T_{noise_{\leq 2}}$).

Figure 8 gives an example of the taxonomy visualisation provided using the proposed SDG-CAP-EXT methodology. The example gives part of the populated SDG taxonomy for Guatemala and SDGs 2, 3 and 4 (zero hunger, good health and well being, and quality education). Green nodes indicate that the indicator will be met (class T), Red nodes that the indicator that will be not met (class N) and black nodes that we do not know whether the indicator will be met because of a lack of data (class *unknown*). The nodes are labelled with the relevant Goal, Target or Indicator ID, or the relevant descriptor, as appropriate.

The proposed SDG-CAP-EXT methodology also provides a geographic visualisations using ALI scores. An example is presented in Figure 9 for North Africa and Central America regions consider with respect to the evaluation presented in this paper. From the figure, it can be seen that Tunisia is predicted to meet 50% of the SDGs on time, while Mexico is predicted to meet only 24% of the SDGs on time. Recall that the ALI score summarises the likelihood of a country achieving its SDGs on time.

Time Series	SDG Time Series Description	Descriptor (D)	Initial Value	Prediction	Result
3.5.2	Alcohol consumption per capita (aged 15 years and older) within a calendar year (litres of pure alcohol)	15+	514	466.24	Unknown
3.3.1	Number of new HIV infections per 1,000 uninfected population, by sex and age (per 1,000 uninfected population)	ALLAGE_MALE	0.09	0.08	T
8.4.2	Domestic material consumption per unit of GDP, by type of raw material (kilograms per constant 2010 United States dollars)	PET	1.36	1.51	F
8.4.2		NMM	1.5	1.67	F
8.4.2		NMC	1.2	1.35	F
8.4.2		CRO	22.1	23.07	F
8.4.2		NMA	25.5	25.96	F
8.4.2		NFO	18.5	19.82	F
2.c.1	Indicator of Food Price Anomalies (IFPA), by type of product	RIC	0.40852	0.69	Unknown
2.a.1	Agriculture share of Government Expenditure (%)		7.5	7.27	Unknown
3.4.1	Number of deaths attributed to non-communicable diseases, by type of disease and sex (number)	BOTHSEX_CAN	1.022735	1.82	F
3.4.1	Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease (probability)	FEMALE_30-70	0.1	0.1	F
3.4.2	Number of deaths attributed to suicide, by sex (number)	BOTHSEX	9.07	5.44	F
3.9.1	Crude death rate attributed to household and ambient air pollution (deaths per 100,000 population)		3.3	1.26	T

Table 7. Fragment of a SDG-CAP-EXT Country Table for Belize showing selected outcomes

6. Conclusion

In this paper, the SDG-CAP-EXT methodology has been presented to predict the attainment of SDGs with respect to specific geographic segments. The hypothesis that the paper sought to address was that better SDG attainment prediction could be obtained if the prediction was conducted using co-related time series within regions rather than individual time series as in the case of previous work. The central challenge was how best to identify such co-related time series; a challenge compounded by the short length of SDG time series and the presence of many missing values in the UN SDG data set. Several different mechanisms were considered, together with four different data imputation methods. The best method was found to be a combination of the others, and the best data imputation method was found to be Spline. Multi-variate LSTMs were used to conduct the forecasting. The proposed methodology was compared with the SDG-CAP methodology from the literature and univariate LSTM forecasting to test the hypothesis. It was found that the hypothesis was correct; better SDG

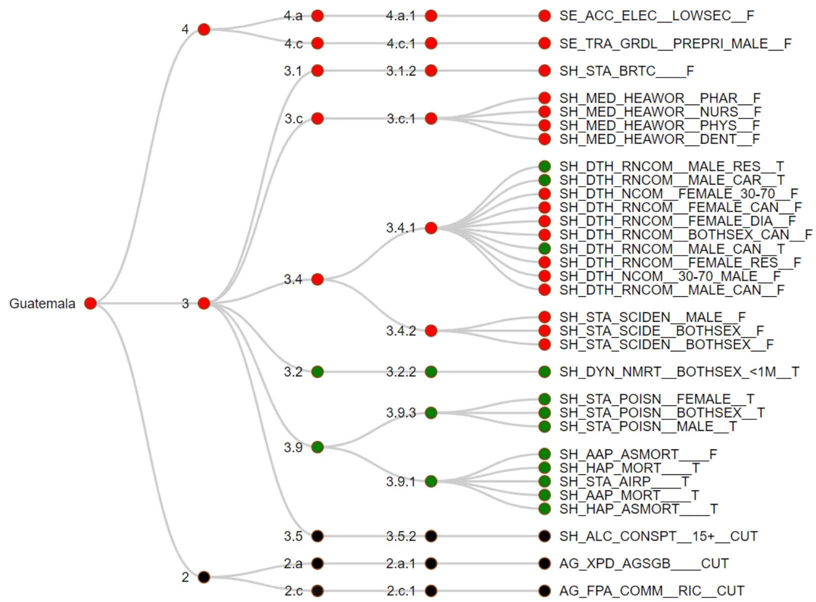

Fig. 8. Visualisation of part of the SDG taxonomy for Guatemala

attainment prediction could be obtained using the SDG-CAP-EXT methodology, which took into consideration co-related time series. It was also demonstrated that the proposed approach was well able to handle short time series.

References

1. Xiangyun Qing and Yugang Niu. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy, 148:461–468, 2018.
2. George Athanasopoulos, Roman A. Ahmed, and Rob J. Hyndman. Hierarchical forecasts for Australian domestic tourism. International Journal of Forecasting, 25(1):146–166, 2009.
3. Kai Chen, Yi Zhou, and Fangyan Dai. A LSTM-based method for stock returns prediction: A case study of China stock market. In Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015, pages 2823–2824. IEEE, 2015.
4. Jan G. De Gooijer and Rob J. Hyndman. 25 Years of Time Series Forecasting. International Journal of Forecasting, 22(3):443–473, 2006.
5. Robert Nau. Notes on the random walk model. Fuqua School of Business, 1(c):1–19, 2014.
6. Rob Hyndman and Andrey Kostenko. Minimum sample size requirements for seasonal forecasting models. Foresight, 6(Spring):12–15, 2007.
7. Yassir Alharbi, Daniel Arribas-Bel, and Frans Coenen. Sustainable development goal attainment prediction: A Hierarchical Framework using Time Series Modelling. IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 1:297–304, 2019.
8. Yassir Alharbi, Frans Coenen, and Daniel Arribas-Bel. Sustainable development goal relational mod- elling: Introducing the SDG-CAP methodology. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 12393 LNCS, pages 183–196, 2020.
9. Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time

Fig. 9. Visualisation of the ALI Scores for the North Africa and Central America Regions

series. In KDD workshop (Vol. 10, No. 16, pp. 359-370).
10. Wold, S., Esbensen, K., Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.
11. suffix Jr suffix William R Kinney. ARIMA and Regression in Analytical Review: An Empirical Test. The Accounting Review, 53(1):48–60, 1978.
12. Sepp Hochreiter and Ju¨rgen Schmidhuber. Long short-term memory.Neural computation, 9(8):1735–1780, 1997.
13. A. J. Lawrance and P. A. W. Lewis. The Exponential Autoregressive-Moving Average Earma(P,Q) Process . Journal of the Royal Statistical Society: Series B (Methodological), 42(2):150–161, 1980.
14. Brownlee Jason. Deep Learning For Time Series Forecasting, volume 1. Machine Learning Mastery, 2018.
15. Paresh Kumar Narayan and Russell Smyth. Multi-variate granger causality between electricity con- sumption, exports and GDP: Evidence from a panel of Middle Eastern countries. Energy Policy, 37(1):229–236, 2009.
16. Halil Altintas and Melike Kum. Multi-variate granger causality between electricity generation, exports, prices and economic growth in Turkey. International Journal of Energy Economics and Policy, 3(SPECIAL ISSUE):41–51, 2013.
17. Hsiao Tien Pao and Chung Ming Tsai. Multivariate Granger causality between CO2 emissions, energy consumption, FDI (foreign direct investment) and GDP (gross domestic product): Evidence from a panel of BRIC (Brazil, Russian Federation, India, and China) countries. Energy, 36(1):685–693, 2011. Wankeun Oh and Kihoon Lee. Causal relationship between energy consumption and GDP revisited: The case of Korea 1970-1999. Energy Economics, 26(1):51–59, 2004.
18. Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference, pages 92–96, 2010.
19. Gyula Do¨rgo, Viktor Sebestye´n, and Ja´nos Abonyi. Evaluating the interconnectedness of the sustain- able development goals based on the causality analysis of sustainability indicators. Sustainability (Switzerland), 10(10):3766, 2018.
20. Meike Nauta, Doina Bucur, and Christin Seifert. Causal Discovery with Attention-Based Convolutional Neural Networks. Machine Learning and Knowledge Extraction, 1(1):312–340, 2019.
21. Bruce B. Frey. Pearson Correlation Coefficient. In The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, pages 1–4. Springer, 2018.
22. Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.
23. Camila Epprecht, D Guegan, and A´ Veiga. Comparing variable selection techniques for linear regres- sion: LASSO and Autometrics. Centre d'e´conomie de la Sorbonne, 2013.
24. Jiahan Li, Weiye Chen. Forecasting macroeconomic time series:LASSO-based approaches and

their forecast combinations with dynamic factor models. International Journal of Forecasting,30(4):996–1015, 2014.

25. Sanjiban Sekhar Roy, Dishant Mittal, Avik Basu, and Ajith Abraham. Stock market forecasting using LASSO linear regression model. In Advances in Intelligent Systems and Computing, volume 334, pages 371–381. Springer, 2015.
26. Shaonan Tian, Yan Yu, and Hui Guo. Variable selection and corporate bankruptcy forecasts. Journal of Banking and Finance, 52:89–100, 2015.
27. Noor Alam and Cynthia Rudin. Robust Nonparametric Testing for Causal Inference in Observational Studies. Optimization Online, Dec, pages 1–39, 2015.
28. Danielsson,P.E.(1980).Euclidean distance mapping.Computer Graphics and image processing,14(3),227-248.
29. Alla Tkachenko, Nataliia Levchenko, Ganna Shyshkanova, Dmytro Plynokos, and Marharyta Kovalenko. Efficiency forecasting for municipal solid waste recycling in the context on sustainable development of economy. In E3S Web of Conferences, volume 166, page 13021. EDP Sciences, 2020.
30. Luis Rivera-Gonza´lez, David Bolonio, Luis F. Mazadiego, and Robert Valencia-Chapi. Long-term electricity supply and demand forecast (2018-2040): A LEAP model application towards a sustainable power generation system in Ecuador. Sustainability (Switzerland), 11(19):5316, 2019.
31. B Hughes, CM Peterson, DS Rothman, and JR Solorzano. Ifs health model documentation,working paper. Denver: Pardee Center for International Futures, 2014.
32. Pierre Horwitz. Millennium development goals, 2018.
33. Statistics Division UN. E-Handbook on Sustainable Development Goals Indicators, 2559.
34. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pretten- hofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
35. Charles A. Hall and W. Weston Meyer. Optimal error bounds for cubic spline interpolation. Journal of Approximation Theory, 16(2):105–122, 1976.
36. Fred T. Krogh. Efficient Algorithms for Polynomial Interpolation and Numerical Differentiation. Mathematics of Computation, 24(109):185, 1970.
37. Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, and Mikko Kolehmainen. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38(18):2895–2907, 2004.
38. Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W. Cottrell. A dual- stage attention-based recurrent neural network for time series prediction. arXiv, 2017.