Sustainable Development Goal Relational Modelling: Introducing the SDG-RMF Methodology[*]

Yassir Alharbi, Daniel Arribas-Bel and Frans Coenen
The University of Liverpool , Liverpool, L69 3BX, UK
{yassir.alharbi,darribas,coenen}@liverpool.ac.uk

Abstract

A mechanism for predicting whether individual regions will meet there UN Sustainability for Development Goals (SDGs) is presented which takes into consideration the potential relationships between time series associated with individual SDGs, unlike previous work where an independence assumption was made. The challenge is in identifying the relationships and then using these relationships to make SDG attainment predictions. To this end, the SDG Relational Multivariate Forecasting (SDG-RMF) attainment prediction methodology is presented. A multivariate forecasting mechanism for forecasting SDGs time series The results demonstrate that by considering the relationships between time series, more accurate SDG forecast predictions can be made.

1 Introduction

Time series forecasting is a significant task undertaken within the context of many application domains. The basic idea is to use the time series past lags to predict single or multiple time steps ahead [Brownlee, 2018]. The complexity of time series increases in the presence of short time series [Rob J *et al.*,] and/or missing values. This paper examines the application of time series analysis to Sustainable Development Goals (SDGs) attainment forecasting [UN, 2015]. The challenge can be summarised as follows: (i) the short time series to be utilised (maximum of 19 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 series description $\rightarrow$ categorical identifier) and (iv) the lack of specific attainment values.

In [Alharbi *et al.*, 2019] a baseline forecast methodology was presented, the SDG Attainment Prediction (SDG-AP) methodology founded on a taxonomic hierarchy, derived from the SDG structure, that was designed to answer questions such as "will a geographical area X meet its goal Y by time T". The methodology in [Alharbi *et al.*, 2019] can be categorised as a "bottom-up single variate hierarchical forecasting" approach. The classifiers associated with leaf nodes were built using the time series available from the UN SDG data set. The remaining nodes in the hierarchy held simple boolean functions. The assumption was that the time series associated with each SDGs were independent of each other, although inspection of the various SDGs indicates that it can be anticipated that many of the time series are correlated in some way. A multivariate approach, as opposed to the univariate approach presented in [Alharbi *et al.*, 2019], therefore seemed appropriate. In [Alharbi *et al.*, 2020] the SDG Correlation/Causal Attainment Prediction (SDG-CAP) methodology was presented that took into consideration the relationships that may exist between SDG time series. The proposed methodology can be categorised as a "bottom-up hierarchical multivariate time series forecasting approach".

Given the foregoing, this paper presents the SDG Relational Multivariate Forecasting (SDG-RMF) methodology. The SDG-RMF methodology is founded on the ideas presented [Alharbi *et al.*, 2019] and [Alharbi *et al.*, 2020] but proposes a hybrid strategy. The proposed methodology can be categorised as a "bottom-up hierarchical univariate/multivariate time series forecasting approach" to forecasting SDG attainment based on historical SDG data. This is achieved using both univariate and multivariate forecasting methods coupled with a filtration mechanism founded on the approach used in [Alharbi *et al.*, 2020] to identify correlated time series.

2 Literature Review

In this section, a review of existing work directed at time series forecasting is presented. Time series forecasting can be expressed either in terms of univariate forecasting or in terms of multivariate forecasting [Gooijer and Hyndman, 2006]. With respect to the specific mechanisms investigated in this paper two mechanisms are considered Fbprophit [Taylor and Letham, 2017] and Multivariate multi-step encoder-decoder LSTM [Brownlee, 2018].

Fbprophet is an additive model proposed by Facebook [Taylor and Letham, 2017]. The model decompose a time series y into three main parts, trend (g), seasonality (s) and holidays (h), plus an error term e, as shown in Equation 1. For the SDG time series, only g is relevant. Fbprophet was

[*] These match the formatting instructions of IJCAI-07. The support of IJCAI, Inc. is acknowledged.

used in [Alharbi *et al.*, 2019] to forecast SDGs attainment and is thus used for comparison purposes later in this paper.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{1}$$

While linear models such as ARMA and ARIMA [Gooijer and Hyndman, 2006] have been widely adopted in, and associated with, time series forecasting; non-linear models, inspired by neural networks, such as LSTMs, have received much attention in the past few years. LSTMs were first introduced in 1997 in [Hochreiter and Schmidhuber, 1997], and have been widely adopted ever since, especially in domains such as weather prediction [Qing and Niu, 2018] and stock market prediction [Chen *et al.*, 2015]. With respect to the evaluation presented later in this paper, both single variate and multivariate LSTMs are considered.

3 The United Nations' Sustainable Development Goal Agenda

At the beginning of the twentieth century, the United Nations (UN) announced its vision for a set of eight development goals, that all member states would seek to achieve [United Nations Development programme, 2007]. These were referred to, for obvious reasons, as the Millennium Development Goals (MDGs). In 2015, the UN extended the initial eight MDGs into seventeen Sustainable Development Goals (SDGs), listed in Table 1, to be achieved by 2030 [Sapkota, 2019; UN, 2015]. Each individual SDG has several target sub-goals, and each target sub-goal has several indicators, sub-indicators and even sub-sub-indicators associated with it; each linked to an attainment threshold of some kind. 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 $g\_s_1\_s_2\_\ldots$, where g is the goal number, s_1 is the sub-goal number, s_2 is the sub-sub-goal number, and so on. For example, SDG 2_22 indicates sub-goal 22 of SDG 2. The UN has made available the MDG/SDG data collated so far[1].

In Alharbi et al. [Alharbi *et al.*, 2019], the complete set of SDGs and associated targets and sub-sub-indicators were conceptualised as a taxonomic hierarchy, as shown in Figure 1. In the figure the root node represents the complete set of SDGs, the next level the seventeen individual SDGs, then the target sub-gaols referred to as "targets", the indictors and sub-indicators and so on. The same taxonomy is used 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 coupled with SDG indicator. Each row is date stamped. Currently, the data set features 283 different geographical regions, and for each region, there are, as of October 2019, up to 801 different time series [Dörg\Ho *et al.*, 2018]. The maximum length of a time series was 19 points, covering 19 year's of observations, although a time series featuring a full 19 observations is 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 form the year 2000 onward was considered; 127,429 time series in total. By applying time series analysis to the data, trends can be identified for prediction/forecasting purposes (see for example [Alharbi *et al.*, 2019]).

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 1: The seventeen Sustainable Development Goals (SDGs)

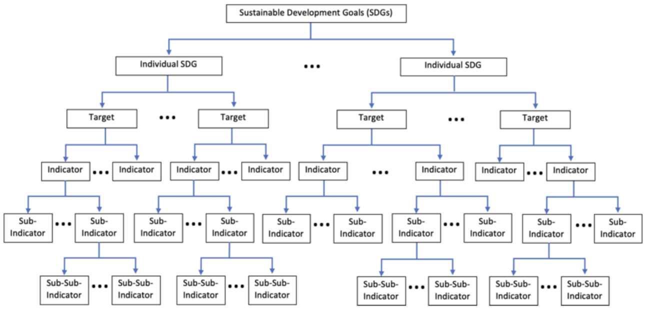

Figure 1: The hierarchical nature of SDGs data

The number of missing values in the SDG data set presented a particular challenge (see Figure 2). The total theoretical number of observations (time series points) in the data was 2,548,580, while the actual number was 1,062,119; in other words, the data featured 1,486,461 missing values (58.3% of the total). Most of these missing values were missing in what can only be described as a random manner. However, in other cases, the missing data could be explained because observations were only made following a five-year cycle.

4 The SDG-RMF Attainment Prediction Methodology

Figure 3 provides an overview of the proposed SDGs-RMF) Methodology. The first step is data preprocessing to transpose the data into an appropriate format for the application of machine learning. The next step is to filter the time series ac-

[1] https://unstats.un.org/SDGs/indicators/database/

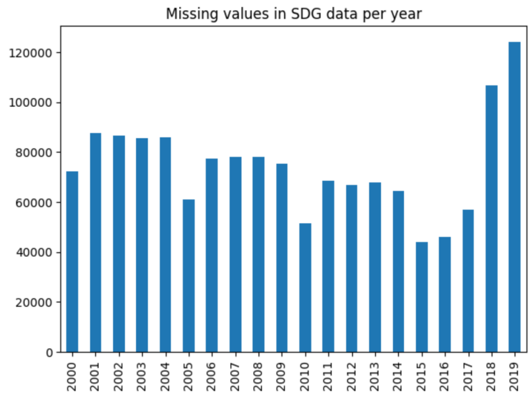

Figure 2: Number of missing values in SDG data set per year.

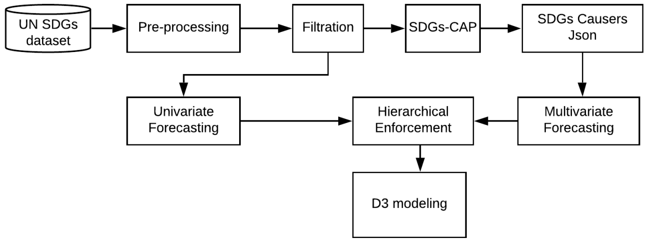

Figure 3: Schematic of the SDGs Releational Multivariate Forecasting (SDGs-RMF) Methodology

cording to the number of missing values in each. An arbitrary number of missing values threshold of 10 was used. Time series with more missing values than the threshold were selected with respect to the univariate forecasting. The remaining time series were used with respect to multivariate times series forecasting. Several statistical and deep learning tests (see [Alharbi *et al.*, 2020]) were considered for finding hidden relationships in the data. The discovered relations were then stored as a JSON file. In the next step ”multivariate forecasting” the top 10 related time series for each time series were selected for the multivariate LSTM. Time series with over 10 missing values were forecast using a univariate approach, specifically the Fbprophet model. In the next step hierarchical enforcement was used to first recombine each lower level indicator under its proper location; then, using the taxonomy obtained in [Alharbi *et al.*, 2019], each node was checked for the predicted value at the targeted date. The last step was to visualise the results using D3 visualisation.

5 Evaluations, Discussion and Visualisation

For the evaluation of the proposed SDG-RMF methodology the two forecasting mechanism listed earlier were used: (i) Fbprophet and (ii) multivariate LSTM. For the evaluation report here the geographic area Egypt was used together with SDG Target 3.2 (SDG 3_2), “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”. Target 3.2, encompasses six indicators, these are listed in Table 2. Recall that earlier in this paper, it was hypothesised that by combining related time series, better predictions could be made compared to results presented in [Alharbi *et al.*, 2019] where an independence assumption was made. To measure this, the SDG-RMF methodology was run, using time series from the geographic area Egypt and SDG 3_2 data points from 2001 to 2013 for training, and 2014-17 for predictions. This was the same data set used in [Alharbi *et al.*, 2019], hence comparisons could be made. RMSE was used as the comparison metric. For the multivariate forecasting, as noted above, the input was limited to the top ten most-related time series; an example is given in Figure 5. In Table 4 it can be seen what exact series codes SDG-CAP methodology found as causers for SD_3_15 the codes could be deciphered from the SDG website [2]. The obtained RMSE results are presented in Table 2. From the table, it can be seen that the multivariate LSTM method produced better forecasting errors than using univariate methods. It is interesting to note; however, that Fbprophet produced better predictions on some occasions. To illustrate the utility of the proposed approach the geographic area Egypt SDG 3_2 were again used. The proposed SDG-RMF methodology was then applied to automatically predict whether the target geographic area would be meet SDG 3_2 by 2030. To predict whether Target 3.2 will be met in 2030, all forecasted values must be less the 25 % of the benchmark value for the year 2015. The results are presented in Table 3. From the table it can be seen that in the case of the geographic area Egypt, Target 3.2 will not be met in 2030.

Indicator	SDG-AP	LSTM
	e_{RMSE}	e_{RMSE}
Infant deaths (number) (Male)	2688	0.47
Infant mortality rate (Per1k) (Male)	0.016	0.50
Under-five deaths (number)	2755	0.62
Under-five mortality rate (Per1k)	0.010	0.41
Neonatal mortality rate (Per1k)	0.016	0.38
Neonatal deaths (number)	66.095	0.29
Average e_{RMSE}	918.18	0.44
Standard Deviation e_{RMSE}	1397.23	0.11

Table 2: RMSE evaluation results using Fbprohphet and Multivariate LSTM

5.1 Framework Visualisation

Each geographical entity has several hundred time series to check; analysing predictions will therefore not be straightforward. Thus D3 [Bostock *et al.*, 2011] visualisation was used to check the attainment in the form of hierarchical dendrograms generated using the D3.js JavaScript library. The prediction visualisation for Target 3.2, with respect to the geographic area of Egypt, is given in Figure 4.

6 Conclusion

In this paper the SDG-RMF methodology has been presented for predicting the attainment of SDGs with respect to specific geographic regions. The hypothesis that the paper sought

[2] https://unstats.un.org/sdgs/metadata/

Goal	Series Description from SDG data	Initial	Prediction	Threshold	Result	Date
3.2.1	Under-five deaths (number) <5Y Male	32890	32156.19	<=25%	Not Met	2030
	Under-five deaths (number) <5Y Female	27542	3074.82	<=25%	Met	2026
	Infant deaths (number) < 1Y Male	28323	29095.0	<=25%	Not Met	2030
	Infant deaths (number) <1Y Female	22967	22329.34	<=25%	Not Met	2030
	Under-five mortality rate (Per 1K) Male <5	25	7.95	<=25%	Met	2023
	Under-five mortality rate (Per 1K) Female <5	22.2	5.39	<=25%	Met	2023
	Infant mortality rate (per 1K) Male <1Y	21.3	10.21	<=25%	Met	2022
	Infant mortality rate (per 1K) Female <1Y	18.7	7.07	<=25%	Met	2029
3.2.2	Neonatal mortality rate (Per 1K) <1M	12.5	5.09	<=12	Met	2030
	Neonatal deaths (number) < 1 M	31796	23308.58	<=25%	Met	2030

Table 3: Framework evaluation using Target 3.2

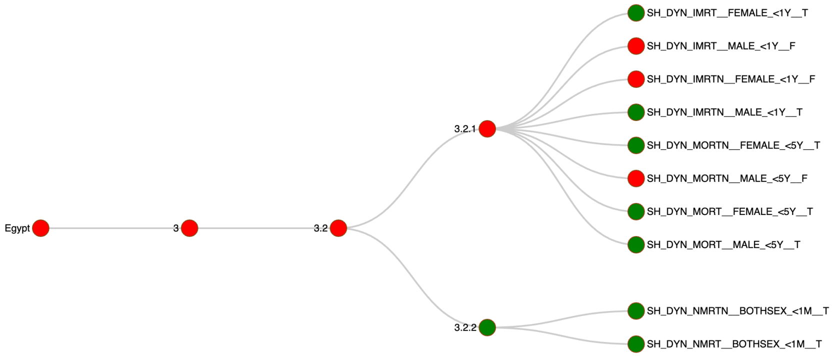
Egypt
3
3.2
3.2.1
3.2.2
SH_DYN_IMRT__FEMALE_<1Y__T
SH_DYN_IMRT__MALE_<1Y__F
SH_DYN_IMRTN__FEMALE_<1Y__F
SH_DYN_IMRTN__MALE_<1Y__T
SH_DYN_MORTN__FEMALE_<5Y__T
SH_DYN_MORTN__MALE_<5Y__F
SH_DYN_MORT__FEMALE_<5Y__T
SH_DYN_MORT__MALE_<5Y__T
SH_DYN_NMRTN__BOTHSEX_<1M__T
SH_DYN_NMRT__BOTHSEX_<1M__T

Figure 4: D3 visualisation for Goal 3 Target 2.1 and 2.2

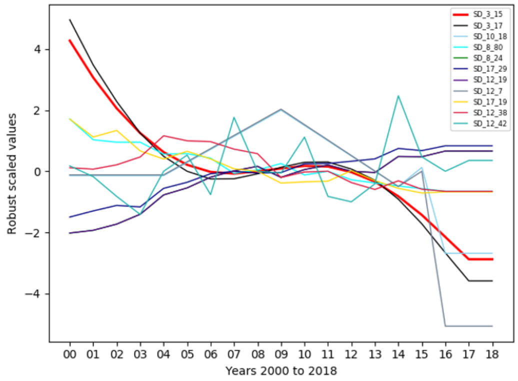

Figure 5: An example of related SDG time series; the example shows ten time series which affect the SDG 3_15 time series for the geographic region Egypt

to address was that better SDG attainment prediction could be obtained if the prediction was conducted using co-related time series rather than individual time series as in the case of previous work. The central challenge was prediction using short length time series, as in the case of the UN SDG data, and the presence of many missing values in the data. Multivariate LSTM were used to conduct the forecasting. To test the hypothesis, the proposed methodology was compared with the SDG-AP methodology from the literature . It was found that the hypothesis was correct, better SDG attainment prediction could be obtained using the SDG-RMF methodology, which took into consideration co-related time series. For future research, the intention is firstly to incorporate the proposed approach to consider co-related time series across geographic regions, not just within a single geographic region as in the case of this paper, bearing in mind the economic and geographical differences between different regions.

Indicators	Series Code
SD_3_15 (the target variable) SD_3_17	SH_DYN_IMRTN
SD_10_18	TM_TRF_ZERO
SD_8_80,SD_8_24,SD_12_38, SD_12_42,SD_12_7	EN_MAT_DOMCMPG
SD_17_29	TM_TAX_ATRFD
SD_12_19	SG_HAZ_CMRMNTRL
SD_17_19	TM_TAX_WWTAV

Table 4: The SDGs series codes found to effect the attainment of Egypt 3.2.1 Male (SD_3_15)

References

[Alharbi *et al.*, 2019] Yassir Alharbi, Daniel Arribas-Be, and Frans Coenen. Sustainable Development Goal Attainment Prediction: A Hierarchical Framework using Time Series Modelling. In *KDIR*, 2019.

[Alharbi *et al.*, 2020] Yassir Alharbi, Daniel Arribas-Be, and Frans Coenen. Sustainable Development Goal Relational

Modelling: Introducing the SDG-CAP Methodology. In *DAWAK*, 2020.

[Bostock *et al.*, 2011] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, dec 2011.

[Brownlee, 2018] Jason Brownlee. *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2018.

[Chen *et al.*, 2015] Kai Chen, Yi Zhou, and Fangyan Dai. A LSTM-based method for stock returns prediction: A case study of China stock market. In *2015 IEEE international conference on big data (big data)*, pages 2823–2824. IEEE, 2015.

[Dörg\Ho *et al.*, 2018] Gyula Dörg\Ho, Viktor Sebestyén, and János Abonyi. Evaluating the Interconnectedness of the Sustainable Development Goals Based on the Causality Analysis of Sustainability Indicators. *Sustainability*, 10(10):3766, 2018.

[Gooijer and Hyndman, 2006] Jan G De Gooijer and Rob J Hyndman. 25 years of time series forecasting. *Int. J. Forecast*, 2006.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Qing and Niu, 2018] Xiangyun Qing and Yugang Niu. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy*, 148:461–468, 2018.

[Rob J *et al.*,] Hyndman Rob J, Athanasopoulos George, and Shang Han Lin. hts: An R Package for Forecasting Hierarchical or Grouped Time Series.

[Sapkota, 2019] Shaswat Sapkota. *E-Handbook on Sustainable Development Goals*. United Nations, 2019.

[Taylor and Letham, 2017] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 2017.

[UN, 2015] UN. Transforming our World: the 2030 Agenda for Sustainable Development. Working papers, eSocialSciences, 2015.

[United Nations Development programme, 2007] United Nations Development programme. Millennium Development Goals, 2007.