Sustainable Development Goal Relational Modelling: Introducing the SDG-CAP Methodology

Yassir Alharbi[1,3], Frans Coenen[1], and Daniel Arribas-Bel[2]

[1] Department of Computer Science,
The University of Liverpool, Liverpool L69 3BX, United Kingdom
[2] Department of Geography and Planning,
The University of Liverpool, Liverpool L69 3BX, United Kingdom
[3] Almahd College, Taibah University
Al-Madinah Al-Munawarah, Saudi Arabia
{yassir.alharbi,coenen,d.arribas-bel}@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 existence of relationships and then using these relationships to make SDG attainment predictions. To this end the SDG Correlation/Causal Attainment Prediction (SDG-CAT) methodology is presented. Five alternative mechanisms for determining time series relationships are considered together with three prediction mechanisms. The results demonstrate that by considering the relationships between time series, by combining a number of popular causal and correlation identification mechanisms, more accurate SDG forecast predictions can be made.

Keywords: Time series correlation and causality, Missing values, Hierarchical classification, Time series forecasting, sustainable development goals.

1 Introduction

Time series forecasting is a significant task undertaken across many domains. The basic idea, given a previously unseen time series, is to predict the next point or points in the series. This is usually conducted using single-variate time series, although in some cases multi-variate time series are considered [4, 23]. Given a short time series this is a particular challenge [12]. One application domain where this is the case is in the context of the data published with respect to the United Nations (UN) Sustainability for Development Goals (SDGs) [25]. Where, at time of writing, data spanning only 19 years was available; in other words times series comprised of a maximum of only 19 points. The SDG short time series challenge is compounded by the large number of missing values that are a feature of the data set, meaning that many time series comprise fewer than 19 points. The aim here is to use the available short time series data to forecast whether a particular geographic region will meet the UN SDGs or not.

In [1] a SDG Attainment Prediction (SDG-AP) methodology was presented founded on the idea of a taxonomic hierarchy and designed to answer the question *"will geographic region x meet goal y by time t"*. The solution was conceptualised as a bottom-up hierarchical Boolean ("yes/no") classification problem. Each node within the taxonomic hierarchy had a Boolean classifier associated with it. The classifiers associated with the leaf nodes were built using the time series available within the UN SDG data set. The remaining nodes in the tree were associated with simple Boolean functions that took input from their child nodes. However, the leaf node classifiers were built assuming that each goal was independent of any other goals. This is clearly not the case. For example, the "No Poverty" and "Quality Education" SDGs are clearly related. Similarly, the time series associated with the goal "Clean Water and Sanitation" in (say) the geographic region "Egypt" are clearly related to the time series associated with the same goal in similar regions.

The hypothesis presented in this paper is that better SDG attainment prediction accuracy can be obtained by considering the possible relationships between SDG time series. Thus, with

reference to [1], instead of building each leaf node classifier according to the relevant time series data (a one-to-one correspondence), it is proposed in this paper that it might be better if the time series data sets used to build the classifiers were more comprehensive, in other words, founded on a set of co-related time series. The challenge is then how to identify these related time series.

Given the above this paper proposes the SDG Correlation/Causal Attainment Prediction (SDG-CAP) methodology designed to address the disadvantages associated with the work presented in [1], although the work in [1] provides an excellent forecasting benchmark. The main challenge is determining which time series are influenced by which other time series. This can be done by hand given a domain expert and sufficient time resource. However, automating the process is clearly much more desirable. The work presented in this paper provides a potential solution to this problem with a focus on time series within the same geographic region, as opposed to the same time series across different geographic areas (the latter is an item for future work). In the context of the proposed SDG-CAP methodology, this paper makes three contributions:

1. An investigation into mechanisms whereby relationships between short time series can be discovered.
2. A comparative investigation of missing value imputation methods.
3. The usage of multi-variate time series forecasting given known relationships across short time series.

Five mechanism are considered whereby relationships between time series can be discovered: (i) Granger Causality [9], (ii) Temporal Causal Discovery Framework (TCDF) [19], (iii) Least Absolute Shrinkage Selector Operator (LASSO) [28], (iv) Pearson Correlation [3] and (v) a combination of all four. The effectiveness of the proposed mechanism is considered by comparing the forecast results produced with those given in [1] using Root Mean Square Error (RMSE) [13] as the comparater metric and a number of forecasting mechanisms.

The rest of the paper 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 methodology is then described in 4 and the evaluation of the proposed methodology in Section 5. The paper concludes with a summary of the main findings, and a number of proposed direction for future research, in Section 6.

2 Literature Review

In this section, a review of existing work directed at discovering relationships between time series is presented. A relationship between two-time series can be expressed either in terms of causality [2] or in terms of correlation [3]. Causality implies that a change in one variable results in a change in the other in either a positive or a negative manner. An alternative phrase for causality is "strong relationship". Correlation implies that the values associated with two variables change in a positive or negative manner with respect to one another [3]. Correlation can be viewed as a specialisation of causality, implying that a causal relationship signals the presence of correlation; however, the reverse statement does not hold. Each is discussed in further detail in the following two sub-sections, Sub-sections 2.1 and 2.2, with respect to the specific mechanisms investigated in this paper. Five mechanisms are considered in total, two causality mechanisms, two correlation mechanisms and a combined mechanism. The first four were selected because they are frequently referenced in the literature. Collectively we refer to these mechanisms as *filtration* methods [33] because they are used to filter time series data (specifically SDG time series) so as to determine which time series are related in some way. This section then goes on, Sub-section 2.3, to consider relevant previous work directed at time series forecasting

2.1 Causality

As noted above, two causality mechanisms are considered in this paper: Granger Causality and the Temporal Causal Discovery Framework (TCDF). Granger causality is the most frequently cited mechanism for establishing causality found in the literature [16, 18, 20, 21]. Granger Causality

was introduced in the late 60s [9] and is fundamentally a statistical test of the hypothesis that an independent time series x can be used to forecast a dependent time series y. Granger causality is determined as defined by Equation 1 using the value of time series t "past lags", and the value of time series y past lags plus a residual error e. Granger Causality has been used previously to determine the relationship between pairs of values in SDG time series as reported in [6]. However, the study was only able to find 20,000 pairs of values in the SDG data that featured causality, out of a total of 127,429 time series. It should be noted that the study only considered time series with ten or more observations, ignoring time series with a proportionally high number of missing values; this may be considered to be a limitation of this study.

$$y_t = a_1 y_{t-1} + b_1 x_{t-1} + e \tag{1}$$

The TCDF is a more recent mechanism than Granger Causality [19]. It is a deep learning framework, founded on the use of Attention-based Convolutional Neural Networks, to discover non-linear causal relationships between time series. TCDF can find "confound delays" where the past time series can trigger a change in not only the next temporal step but in a number of future steps. TCDF is considered to be the current state of the art mechanism for causality discovery in time series because it outperforms many previously proposed mechanisms, including Granger Causality. One major limitation of TCDF, at least in the context of SDG time series, is that it does not perform well on short time series.

2.2 Correlation

Two correlation mechanism are considered in this paper: Pearson Correlation and the Least Absolute Shrinkage Selector Operator (LASSO). Pearson Correlation is one of the most frequently used correlation tests [3]. The Pearson Correlation coefficient is a number between $+1$ and -1 and shows how two variables are linearly related. It has been used in many studies to determine the nature of the linearity between variables [22, 5, 31]. The basic formula for Pearson is given in Equation 2 where n is the number of observations, y_i is a value in time series Y, and x_i is a value in time series X.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{2}$$

Lasso [28] is a regression analysis method frequently used when respect to high dimensionality time series data; data featuring many variables, some of which may not be relevant. It is another widely used method [7, 17, 24, 27]. LASSO reduces the dimensionality by penalising variances to zero, which will remove irrelevant variables from the model. From inspection of Equation 3 it can be seen that the first part is the normal regression equation. The second part is a penalty applied to individual coefficients. If λ is equal to 0, then the function becomes a normal regression. However, if λ is not 0 coefficients are penalised.

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3}$$

2.3 Time Series Forecasting

Three time series forecasting mechanisms are considered in this paper: (i) Fbprophet, (ii) Multivariate Long short-term memory (LSTM) and (iii) Univariate LSTM. Fbprophet is an additive model proposed by Facebook [26]. 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 4. For the SDG time series only g is relevant. Fbprophet was used in [1] to forecast SDGs attainment and is thus used for comparison purposes later in this paper.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{4}$$

While linear models such as ARMA and ARIMA [8] have been widely adopted in, and associated with, time series forecasting; non-linear models, inspired by neural networks, such as LSTM, have received a lot of attention in the past few years. LSTM were first introduced in 1997 in [11], and have been widely adopted ever since, especially in domains such as weather predictions [23] and stock market predictions [4]. With respect to evaluation presented later in this paper both single variate and multivariate LSTM are considered.

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)

3 The United Nations' Sustainable Development Goal Agenda

In 2000 the United Nations (UN) announced its vision for a set of eight development goals, listed in Table 1, that all member states would seek to achieve [30]. 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 2, to be achieved by 2030 [25, 29]. Each SDG has a number of sub-goals and sub-sub-goals 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].

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)

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

In Alharbi et al. [1] the complete set of SDGs and associated sub- and sub-sub-goals was 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 sub-goals referred to as "targets", the sub-sub-gaols referred to as "indicators" and so on. The same taxonomy is used with respect to the work presented in this paper.

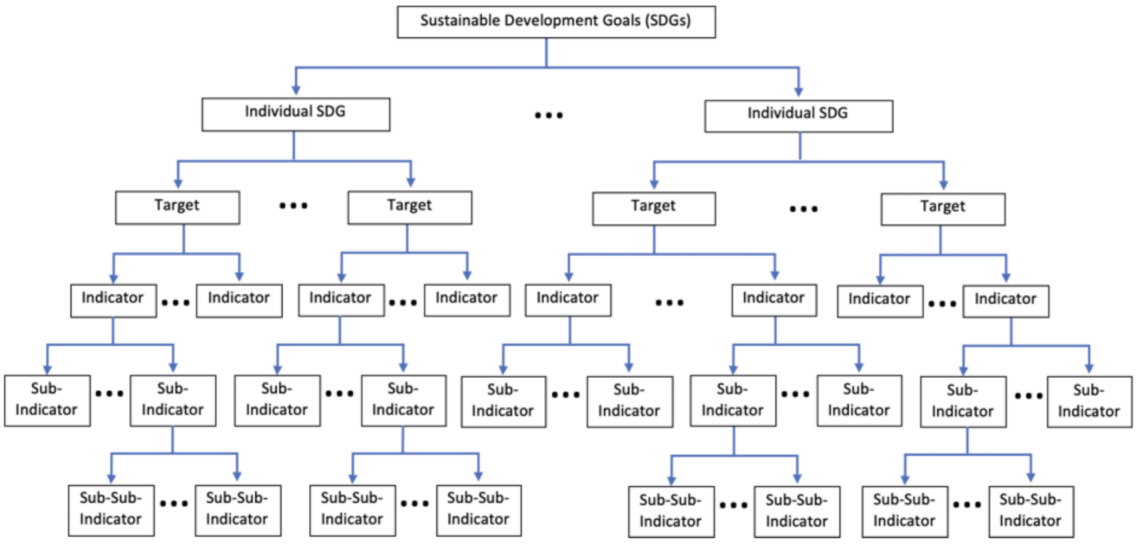


Fig. 1. The hierarchical nature of SDGs data [1]

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 sub-goals and sub-sub-goals. Each row is date stamped. The data set features 283 different geographical regions, and for each region there are, as of October 2019, up to 801 different time series [6]. 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 was 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 [1]).

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, but in other cases, the missing data could be explained because observations were only made following a five-year cycle.

4 The SDG Correlated/Causal Attainment Prediction Methodology

A schematic of the proposed SDG Correlated/Causal Attainment Prediction (SDG-CAP) Methodology is presented in Figure 3. The input is the collection of SDG time series associated with a geographic region of interest. The output is a attainment prediction model. The input data is preprocessed in three steps: (i) Flatning, (ii) Imputing and (iii) Rescaling. During flatning [32] the input time time series were reshaped so that every record comprised a tuple of the form: $\langle Country, Goal, Target, Indicator, CategoricalIdentifiers, \{v_{2000}, v_{2001}, \ldots, v_{2019}\}\rangle$, where v_i is a value for the year i, and the categorical identifiers are things like gender and/or age which may be relevant to a particular goal. It was noted earlier that the SDG data collection features many missing values. For the purposes of the work presented in this paper, any time series with less

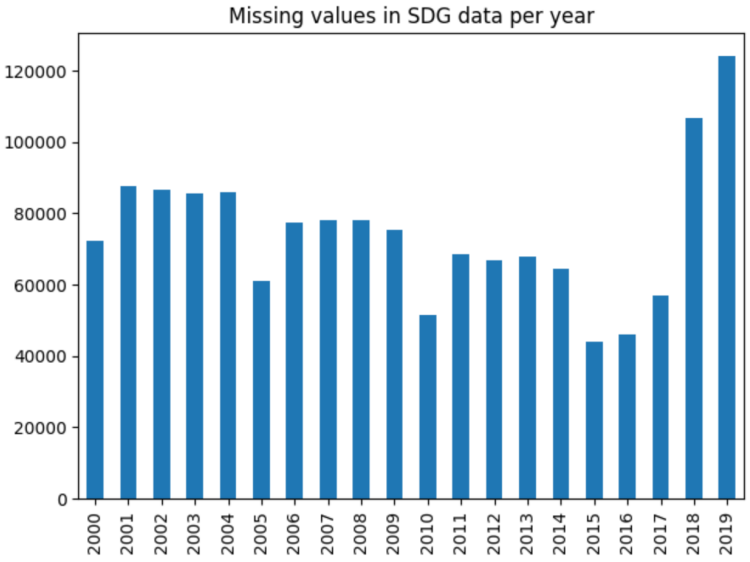


Fig. 2. Number of missing values in SDG data set per year.

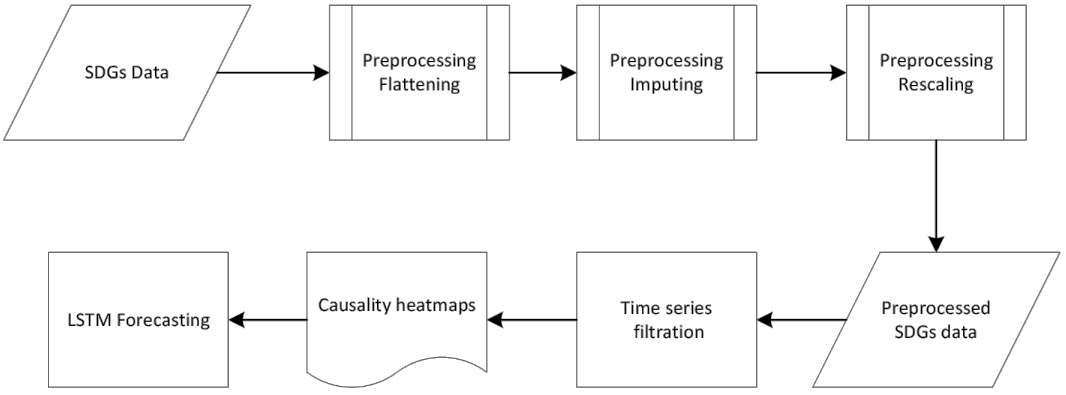


Fig. 3. Schematic of the SDG Causality/Correlation Attainment Predict (SDG-CAP) Methodology

than five values was removed during the flatning stage. As a consequence, the total number of time series to be considered was reduced from 127,429 to 80,936. However, the remaining time series still featured up to thirteen missing values. The next step in the preprocessing was therefore to impute missing values. Experiments were conducted using four different imputation methods: (i) Linear, (ii) Krogh, (iii) Spline and (iv) Pchip [10, 15, 14]. The aim was to identify the most appropriate imputation method. The experiments were conducted using complete time series only from Egypt target 3.2 time series. A 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 3 with respect to Egypt target 3.2. 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 SDG-CAP. The final preprocessing step was to rescale the time series data so that it was referenced to a uniform scale. The reason for this is that the values in the various time series are referenced too many different numeric ranges, for example population counts in the millions against observed values in single digits.

Once a "clean" data set had been generated the next step was to identify relationships within the data using an appropriate filtration method. As noted earlier experiments were conducted, reported

SDG3.2.	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 3. RMSE comparison of imputation methods used to generate missing values in the SDG data (best results in bold font)

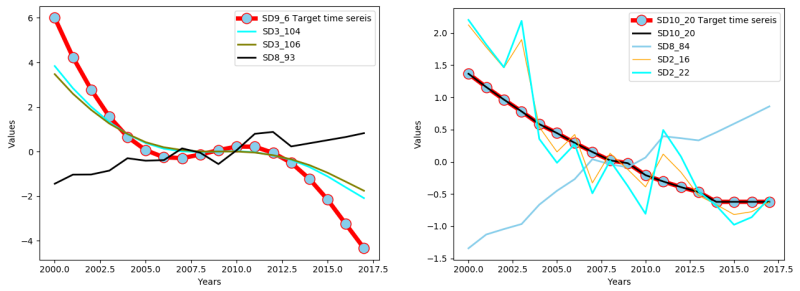
on in the following evaluation section, using five different filtration mechanisms. Regardless of which filtration method is used, the outcome was presented as a heat map. An example fragment of a heat map generated using LASSO, and the geographic area Egypt, is given in Figure 4. The darker the colour the greater the LASSO R^2 value. The leading diagonal represents SDG comparisons with themselves; hence, as expected, these are highly correlated. The heat map was then used to collate time series for the purpose of prediction model generation, by using the top 5 highest R^2 values. The number of selected time series with respect to each goal was limited to a maximum of the top five most-related. These groups of time series were then used as the input to a Multivariate LSTM to predict future values which in turn could be used for attainment prediction.

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

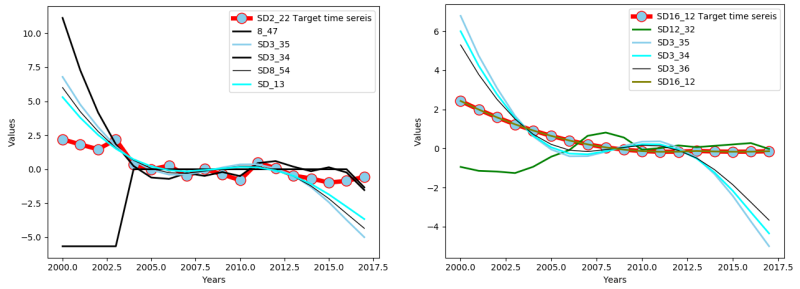
Fig. 4. A fragment of the heat map produced using LASSO analysis and the geographic area Egypt

5 Evaluations

For the evaluation of the proposed SDG-CAT methodology the five different filtration mechanism listed earlier were used. Namely two causality mechanisms (Granger Causality and TCDF), two correlation mechanisms (LASSO and Pearson) and a combination of all four. For the last method the individual results were combined. The evaluation was conducted using the geographic area



SD9_6 SD10_20



SD2_22 SD16_12

Fig. 5. Examples of related time series with respect to particular target SDGs using the combined method

SDG 17_56	Lasso	Granger Causality	Pearson Corr.	TCDF	Combined
SDG 8_13	0.263713282	0.614107986	0.529714622	2	3.407535890
SDG 16_43	0.008389365	0.071975083	0.019141897	2	2.099506345
SDG 9_5	0.000000000	0.941067933	0.000000000	1	1.941067933
SDG 3_35	0.094544491	0.637227132	1.114146883	0	1.845918506
SDG 3_34	0.083353451	0.659473485	0.888387391	0	1.631214327

Table 4. Example time series relationship scores, generated using a range of filtration mechanisms, for SDG 17_56

Egypt. Inspection of the time series for this region, 335 of them, indicated that in many cases there was no data for the years 2018 and 2019; hence seventeen point time series were used covering the time period 2001 to 2017. Some examples of discovered relationships between time series, identified using the combined method, are given in Figure 5. To give one detailed illustration, Table 4 gives the individual relationship scores, using all five mechanisms, for SDG 17_56.

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 [1] where an independence assumption was made. To measure this the SDG-CAT methodology was run with time series for 2001 to 2013, and predictions made for 2014 to 2017. The results given [1] were generated for the same geographic region, Egypt, using Fbprophet. For the multivariate time series LSTM were used, single variate LSTM were also applied to the cleaned data for comparison purposes. RMSE was used as the comparison metric. For the multivariate forecasting, as noted above, the input was limited to the top five most-related time series. The results are presented in Table 5, with best results highlighted in bold font. The method presented in [1] is described as the SDG-AP (SDG Attainment Prediction) method. From the table it can be seen that the combination method produced the best result. All the relation identification mechanisms, coupled with multivariate LSTM, produced better results than the Fbprophet results from [1] and univariate LSTM. It is interesting to note, however, Fbprophet produced better predictions than when using univariate LSTM. It can also be observed, overall, that the proposed SDG-CAP methodology is well able to handle short time series.

SDG	SDG-CAP, Multivariate LSTM					SDG-AP	Univaritae
	Lasso	Grainger	Correlation	TCDF	Combined	Fbprophet	LSTM
2_22	0.0000052	0.0052015	0.0000002	**0.0000002**	0.0000005	0.3231084	1.03190567
3_39	0.0000085	**0.0000001**	**0.0000001**	**0.0000001**	**0.0000001**	0.3180739	0.30534206
6_17	**0.0000002**	0.0000013	0.0000410	0.0000016	0.0000376	0.0420172	0.06473296
8_8	0.0000085	0.0000007	**0.0000006**	0.0000042	0.0000023	0.5924453	4.12603573
9_6	0.0002040	0.0008483	0.0000004	**0.0000002**	**0.0000002**	0.2308795	0.15737508
10_20	0.0000292	0.0000069	0.0047152	**0.0000001**	0.0000002	0.1747101	0.13596105
11_27	0.0000009	0.0000005	0.0000006	**0.0000003**	0.0000008	4.6060315	6.69974063
12_4	0.0000601	0.0000611	0.0000002	**0.0000001**	0.0001074	0.4021510	0.68321411
12_28	0.0000002	0.0000106	**0.0000001**	0.0004241	**0.0000001**	0.5756070	0.33681492
14_4	0.0000622	0.0000002	0.0000178	**0.0000001**	**0.0000001**	0.4025397	0.03858285
15_21	**0.0000003**	0.0022758	0.0000004	0.0085000	0.0000010	1.2691464	1.59389898
16_12	0.0000571	**0.0000001**	0.0000004	0.0000281	0.0000006	0.3846588	0.07288927
17_56	0.0000001	0.0000337	**0.0000001**	0.0000002	**0.0000001**	0.5947819	1.86160687
Ave. RMSE	0.0000336	0.0006493	0.0003675	0.0006892	**0.0000116**	0.7627808	1.31600770
Stand. Dev.	0.0000545	0.0014538	0.0012551	0.0022575	**0.0000293**	1.1456063	1.9764099

Table 5. Example RMSE results produced using SDG-CAP with a range of filtration methods, SDG-AP and Univariate LSTM (best results in bold font)

6 Conclusion

In this paper the SDG-CAP methodology has been presented for predicting the attainment of SDGs with respect to specific geographic regions. 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 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. Five different filtration mechanisms were considered, together with four different data imputation methods. The best filtration method was found to be a combination of the four others, and the best data imputation method was found to be Spline. 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 and univariate LSTM forecasting. It was found that the hypothesis was correct, better SDG attainment prediction could be obtained using the SDG-CAP 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. For future research, the intention is firstly to incorporate the proposed SDG-CAP methodology into a hierarchical bottom-up time series forecasting approach of the form presented in [1]. Secondly the intention is 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.

References

1. Yassir Alharbi, Daniel Arribas-Be, and Frans Coenen. Sustainable Development Goal Attainment Prediction: A Hierarchical Framework using Time Series Modelling. In *KDIR*, 2019.
2. Zhang Ben-gong, Li Weibo, Shi Yazhou, Liu Xiaoping, and Chen Luonan. Detecting causality from short time-series data based on prediction of topologically equivalent attractors. *BMC Systems Biology*, 11:141–150, 2017.
3. Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.
4. 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.
5. K Hema Divya and V Rama Devi. A Study on predictors of GDP: Early Signals. *Procedia Economics and Finance*, 11:375–382, 2014.

6. 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.
7. Camila Epprecht, Dominique Guegan, Álvaro Veiga, and Others. *Comparing variable selection techniques for linear regression: Lasso and autometrics*. Centre d'économie de la Sorbonne, 2013.
8. Jan G De Gooijer and Rob J Hyndman. 25 years of time series forecasting. *Int. J. Forecast*, 2006.
9. Clive W J Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.
10. Charles A Hall and W Weston Meyer. Optimal error bounds for cubic spline interpolation. *Journal of Approximation Theory*, 16(2):105–122, 1976.
11. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
12. Rob J Hyndman. Forecasting: principles and practice, may 2018.
13. Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *Int. J. Forecast*, 2006.
14. 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.
15. Fred T Krogh. Efficient algorithms for polynomial interpolation and numerical differentiation. *Mathematics of Computation*, 24(109):185–190, 1970.
16. Hooi Hooi Lean and Russell Smyth. Multivariate Granger causality between electricity generation, exports, prices and GDP in Malaysia. *Energy*, 35(9):3640–3648, 2010.
17. Jiahan Li and 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.
18. Paresh Kumar Narayan and Russell Smyth. Multivariate Granger causality between electricity consumption, exports and GDP: evidence from a panel of Middle Eastern countries. *Energy Policy*, 37(1):229–236, 2009.
19. 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.
20. 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.
21. 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.
22. Jorge V Pérez-Rodr\'\iguez, Francisco Ledesma-Rodr\'\iguez, and Mar\'\ia Santana-Gallego. Testing dependence between GDP and tourism's growth rates. *Tourism Management*, 48:268–282, 2015.
23. Xiangyun Qing and Yugang Niu. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy*, 148:461–468, 2018.
24. Sanjiban Sekhar Roy, Dishant Mittal, Avik Basu, and Ajith Abraham. Stock market forecasting using LASSO linear regression model. In *Afro-European Conference for Industrial Advancement*, pages 371–381. Springer, 2015.
25. Shaswat Sapkota. *E-Handbook on Sustainable Development Goals*. United Nations, 2019.
26. Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 2017.
27. Shaonan Tian, Yan Yu, and Hui Guo. Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100, 2015.
28. Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.
29. UN. Transforming our World: the 2030 Agenda for Sustainable Development. Working papers, eSocialSciences, 2015.
30. United Nations Development programme. Millennium Development Goals, 2007.
31. Peter Vinkler. Correlation between the structure of scientific research, scientometric indicators and GDP in EU and non-EU countries. *Scientometrics*, 74(2):237–254, 2007.
32. Earo Wang, Dianne Cook, and Rob J Hyndman. A new tidy data structure to support exploration and modeling of temporal data. *arXiv e-prints*, page arXiv:1901.10257, jan 2019.
33. Xiangzhou Zhang, Yong Hu, Kang Xie, Shouyang Wang, E W T Ngai, and Mei Liu. A causal feature selection algorithm for stock prediction modeling. *Neurocomputing*, 142:48–59, 2014.