DOI: 10.1590/1413-812320182311.26142016

Forecasting the human development index and life expectancy in Latin American countries using data mining techniques

Celso Bilynkievycz dos Santos [1]
Luiz Alberto Pilatti [2]
Bruno Pedroso [1]
Deborah Ribeiro Carvalho [3]
Alaine Margarete Guimarães [1]

Abstract *The predictability of epidemiological indicators can help estimate dependent variables, assist in decision-making to support public policies, and explain the scenarios experienced by different countries worldwide. This study aimed to forecast the Human Development Index (HDI) and life expectancy (LE) for Latin American countries for the period of 2015-2020 using data mining techniques. All stages of the process of knowledge discovery in databases were covered. The SMOReg data mining algorithm was used in the models with multivariate time series to make predictions; this algorithm performed the best in the tests developed during the evaluation period. The average HDI and LE for Latin American countries showed an increasing trend in the period evaluated, corresponding to 4.99 ± 3.90% and 2.65 ± 0.06 years, respectively. Multivariate models allow for a greater evaluation of algorithms, thus increasing their accuracy. Data mining techniques have a better predictive quality relative to the most popular technique, Autoregressive Integrated Moving Average (ARIMA). In addition, the predictions suggest that there will be a higher increase in the mean HDI and LE for Latin American countries compared to the mean values for the rest of the world.*

Key word *Forecasting, Data mining, HDI, Life expectancy, Latin America*

[1] Setor de Ciências Biológicas e da Saúde, Universidade Estadual de Ponta Grossa. Av. General Carlos Cavalcanti 4748, Uvaranas. 84030-900 Ponta Grossa PR Brasil. bilynkievycz@uepg.br
[2] Universidade Tecnológica Federal do Paraná. Curitiba PR Brasil.
[3] Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná. Curitiba PR Brasil.

Introduction

Most Latin American countries are undergoing very similar human development processes, possibly due to the historical context of their political emancipation and their social and cultural characteristics. This development process can be evaluated using the Human Development Index (HDI) adopted by the United Nations Development Programme (UNDP) to measure the progress in a country's quality of life[1-3] based on the geometric mean of education, health and income indicators[4], which classifies most Latin American countries as developing countries with a high HDI[5].

Because it includes a "health" component, measured by a long and healthy life, the index is widely used in health research[6-13], and its sub-component most used in this type of studies is life expectancy (LE)[14-27], which is also among the indicators most used to assess the socioeconomic development of a country.

A very large number of recent studies[6-27] use LE or the HDI as guiding variables in health studies.

The predictability of the HDI or its components can help with government decision-making in terms of whether or not to support public policies if the actual figures match the forecasts. The forecasts can also be used in prospective studies in different fields, including health, to explain the future behaviour of dependent variables.

The literature provides several forecasting techniques, such as forecasting using data mining (DM) techniques applied in different fields[28-34], including health[33,34]. However, no studies were found predicting the HDI or LE for Latin American countries.

Given this gap in the literature, the present study aims to predict the HDI and LE for Latin American countries for the 2015-2020 period, based on historical data and using DM techniques.

This study seeks to contribute to the forecasting of these indicators used in epidemiological research and to the evaluation of the algorithms and models used, based on comparisons between the forecasts and the trends reported by the UNDP with regard to the HDI for periods prior to the forecasts and between forecast quality measures.

Materials and methods

Based on the historical HDI data of 188 countries affiliated with the UNDP, covering the period from 1990 to 2014, all stages of the Knowledge Discovery in Databases (KDD) process were covered[35] and are presented in the subsections below. This process helped determine the algorithm and model that best predicted the HDI and LE for the 22 Latin American countries affiliated with the UNDP for the 2015-2020 period.

Most of the stages of the KDD process were performed in the DM environment powered by WEKA[36], using the Forecast technique, in the application programming interface version 3.7 or later.

During the process, the performance of different function-based algorithms was evaluated. Using the best-performing algorithm, forecasting models were developed, and their results were compared to the latest UNDP reports to identify the most efficient models.

The following measures of quality of the time series predictions were used to evaluate the results: mean absolute error (MAE), mean square error (MSE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), directional accuracy (DAC), relative absolute error (RAE), and root relative squared error (RRSE). In addition, statistical tests of analysis of variance and Student's t-test paired by country were used at different stages of the KDD process, adopting alpha = 0.05 as the significance level.

In parallel to the KDD, a forecast model was developed using the SPSS software and the most popular forecasting technique, Autoregressive Integrated Moving Average (ARIMA), to compare the forecasts to those obtained using the DM techniques at the end of the tests.

Data mining pre-processing

The first step in the pre-processing stage was obtaining the HDI and LE data from the UNDP's database[37], updated on July 24, 2014, and from its 2013 Human Development Report[3]. This data source may be updated at any time, and whenever a new annual Human Development report is released, the time series may undergo more significant updates.

Using these sources, a specific database was developed in Microsoft Access that comprised the time series referring to the period from 1980 to 2013. After the implementation of this database, the KDD stage of "database exploration" was per-

Ciência & Saúde Coletiva, 23(11):3745-3756, 2018

formed using Structured Query Language (SQL), resulting in the descriptive statistics of the time series (Tables 3 and 4).

At the end of the DM pre-processing stage, 90 HDI time series were selected for testing and separated into two batches of data, the first one to test the predictions of the HDI 2013 and the second to test the predictions of the HDI 2014, using data prior to the forecast period. Each batch of data was used for the development of the following: i) a global multivariate model (GMM) trained with multivariate series corresponding to the 188 countries affiliated with the UNDP; ii) 22 specific multivariate models (SMM) trained with groups varying from two to 45 countries, with explanatory power for the index of each Latin American country; and iii) 22 univariate models (UM) trained with series corresponding to each Latin American country, resulting in a total of 45 models per data batch. The GMM was trained with data from 188 countries to increase the algorithm's learning.

For the development of SMMs, HDI datasets for candidate countries of predictors from each Latin American country (target attribute) were selected. The datasets were chosen using the Correlation-based Feature Selection (CFS) algorithm[38], using the cross-validation method. This algorithm prioritizes sets of attributes (independent variables) that are closely related to the target attribute (dependent variable) and weakly related to each other.

Data mining

In this stage, the algorithm most suitable for the study was selected, testing the function-based learning algorithms Least Median Squared, Linear Regression, Multilayer Perceptron, RBF Network, SMOReg, and Gaussian Processes.

To reduce operational costs, the preliminary tests were performed only for the HDI 2013, for which the *SMOReg* algorithm[39] was selected because it generated the best results for the different model categories (Table 1).

At the end of the DM stage, 90 models were developed for completing the tests using the *SMOReg* algorithm: two GMMs, 44 SMMs and 44 UMs. These models were compared with the HDI forecast of 22 Latin American countries for the period 2013-2014 to choose the best performing model (Table 2).

Data mining post-processing

The model results were entered in a database that allowed comparisons between actual values and forecasts as well as between model quality measures. The actual figures for the 2013 HDI were obtained from the UNDP[3] on July 24, 2014, while the figures for the 2014 HDI were consulted after the update was released by the UNDP[5], on December 14, 2015.

After completing all KDD stages for testing algorithms and models, the process was repeated for forecasting the HDI and LE in the 2015-2020 period, applying only the algorithm and the model with the best performance, namely, SMOReg and GMM, respectively. Prior to the forecasting, the 1980-2014 time series were updated on December 14, 2015, since, with the release of each new report, the UNDP database[37] may undergo significant updates[3].

Results

Table 1 presents MAE summary statistics for the tests performed to select the best performing algorithm, which was the SMOReg model.

Table 1. MAE of forecast models developed with the function-based DM algorithm.

	Forecast	2013 HDI					
	Model	GMM		SMM		UM	
	Statistic	μ	±	μ	±	μ	±
Algorithm	SMOReg	0.0002[a]	0.00005	0.0008[to]	0.0005	0.0014[a]	0.0007
	Gaussian Processes	0.0011[b]	0.0008	0.0117[c]	0.0057	0.0174[f]	0.0088
	RBF Network	0.0165[d]	0.0079	0.0161[d]	0.0062	0.0160[e]	0.0070
	Multilayer Perceptron	**	**	0.0021[b]	0.005	0.0020[b]	0.0007
	Linear Regression	**	**	*	*	0.0025[c]	0.0028
	Least Median Squared	**	**	*	*	0.0044[d]	0.0046

$p < 0.05$ in a < b < c < d < e < f (compared by column). * Did not allow testing using sets of countries. ** Did not allow testing using all countries.

Table 2 presents summary statistics of the quality measures of the HDI 2013-2014 forecast for the Latin American countries for selecting the best forecast model.

Table 2 shows that the GMMs exhibited the best forecast quality measures, corresponding to the highest values of DAC and lowest errors (MAE, RMSE, MAPE, RAE and RRSE), compared to SMMs and UMs.

Figure 1 presents the MAE of tests performed with the forecast models adopted in this study and compared with the ARIMA method.

It was observed that the models developed through DM techniques presented the smallest absolute errors relative to the ARIMA model.

Table 3 presents the last five observation points of the historical HDI time series in Latin American countries[37], the index values (2015 to 2020) forecasted by the SMOReg algorithm in GMMs, the statistical summary of the index at the global and Latin America levels and its percentage growth for the forecast horizon.

Figure 2 shows directions and forecasts of the models (dashed lines), trends (continuous lines) reported by the UNDP[3,5] and the 2015-2020 HDI forecast by the GMMs for some Latin American countries that presented the best and worst forecast quality measures, despite not presenting significant differences between the nominal values of the forecasts and the values of the trends already published (2014 and 2015).

Figure 3 shows the global HDI growth curve, with the mean, maximum, minimum and variance values recorded over the period, in addition to the mean for Latin America and Latin American countries with the highest and lowest HDI.

Table 4 presents the last five observation points of the historical LE time series in Latin American countries[37], the values for the variable (2015 to 2020) forecasted by the SMOReg algorithm in GMMs, the statistical summary of the variable at the global and Latin America levels and its percentage growth for the forecast horizon.

Figure 4 presents the global LE growth curve, with the mean, maximum, minimum and variance values recorded over the period, in addition to the mean for developed countries, Latin America and Latin American countries with the highest and lowest LE.

Discussion

HDI forecasts

Regarding the HDI forecasts, it should be noted that significant updates of the indices of some countries may be a limitation of the study. According to the UNDP[1], international and national data estimates may be inconsistent, as international data agencies consult national data and, where appropriate, estimate missing data for inter-country comparisons. In regard to these updates, there were significant differences between HDI values released on July 24, 2014[3], and on December 14, 2015[5].

Some of this study's forecasts, obtained with the forecast model and algorithm selection tests, found HDI trends with a different direction from that found in other studies[3,5] for all the models for Cuba in 2013 (Figure 2a), and Venezuela in 2014 (Figure 2b). Cuba also presented the highest MAE for the 2014 HDI forecast as well as the largest differences between UNDP reports[3,5]. Forecasts for Nicaragua in 2013 (Figure 2c) and

Table 2. Quality measures of the models developed to test the forecast of the HDI in the Latin American countries affiliated with the UNDP.

	Forecast horizon	HDI 2013-2014					
	Model	GMM[a]		SMM[b]		MU[c]	
	Statistic	μ	±	μ	±	μ	±
Quality Measure	Directional accuracy - DAC **	98.61	4.23	96.46	5.54	96.11	5.72
	Mean absolute error - MAE *	0.0002	0.00005	0.0008	0.0005	0.0014	0,0007
	Mean absolute percentage error - MAPE *	0.026	0.006	0.12	0.07	0.21	0.08
	Mean square error - MSE	0.0	0.0	0.0	0.0	0.0	0.0
	Relative absolute error - RAE *	3.53	0.57	17.49	11.28	29.58	15.11
	Root-mean-square error - RMSE *	0.0002	0.00005	0.001	0.001	0.002	0.001
	Root relative squared error - RRSE *	3.51	0.65	25,15	13,80	40,20	17.38

* $p < 0.01$ in a < b < c. ** $p < 0.05$ in a > b > c.

El Salvador in 2014 (Figure 2d) showed the lowest MAE. These comparisons, when forecasts are favourable, as in the case of Nicaragua and El Salvador, can support public and economic policies adopted by these countries to develop the index and, when unfavourable, as in the case of Venezuela, might raise concerns about policies or the data, which may be outdated or inconsistent, as was the case for Cuba.

Bolivia (+0.61%) and Cuba (+0.13%) presented, respectively, the highest and lowest percentage gain in the HDI in the last period (2013-2014) among the Latin American countries, while Venezuela (-0.61%) presented a loss in the index[3,5].

The forecasts obtained in this study (Table 3) show that Uruguay could reach, by 2016, the same level of development as Argentina and

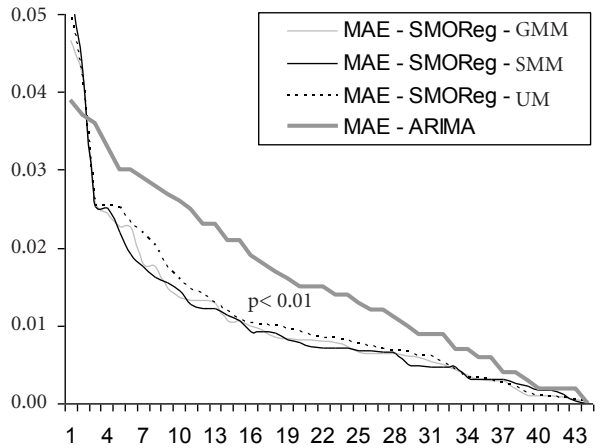

Figure 1. Cumulative MAE per model, resulting from the 2013-2014 HDI forecast for Latin American countries.

Table 3. Latest observation points of the historical HDI series for Latin American countries, its forecasts for 2015-2020 and statistical summary of the index at the global and Latin American levels.

Country*		Latest Observation Points					Forecast Horizon						% Variation
		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2015-2020
ARG		0.811	0.818	0.831	0.833	0.836	0.843	0.85	0.858	0.865	0.874	0.882	5.50
CHL		0.814	0.821	0.827	0.83	0.832	0.836	0.844	0.852	0.86	0.869	0.877	5.43
URY		0.78	0.784	0.788	0.79	0.793	0.799	0.805	0.812	0.819	0.826	0.833	4.98
PAN		0.761	0.759	0.772	0.777	0.78	0.786	0.793	0.8	0.807	0.815	0.823	5.49
CUB		0.778	0.776	0.772	0.768	0.769	0.774	0.78	0.787	0.794	0.802	0.809	5.21
CRI		0.75	0.756	0.761	0.764	0.766	0.77	0.776	0.783	0.789	0.796	0.803	4.78
VEN		0.757	0.761	0.764	0.764	0.762	0.761	0.765	0.769	0.774	0.779	0.785	2.97
MEX		0.746	0.748	0.754	0.755	0.756	0.76	0.766	0.772	0.778	0.784	0.791	4.63
BRA		0.737	0.742	0.746	0.752	0.755	0.761	0.769	0.778	0.786	0.795	0.804	6.45
PER		0.718	0.722	0.728	0.732	0.734	0.739	0.745	0.752	0.759	0.766	0.773	5.26
ECU		0.717	0.723	0.727	0.73	0.732	0.734	0.739	0.743	0.748	0.754	0.759	3.68
COL		0.706	0.713	0.715	0.718	0.72	0.724	0.73	0.736	0.743	0.75	0.757	5.19
BLZ		0.709	0.711	0.716	0.715	0.715	0.715	0.717	0.72	0.723	0.726	0.728	1.86
DOM		0.701	0.704	0.708	0.711	0.715	0.72	0.726	0.733	0.74	0.747	0.754	5.41
PRY		0.668	0.671	0.669	0.677	0.679	0.681	0.686	0.691	0.696	0.701	0.707	4.05
SLV		0.653	0.658	0.662	0.664	0.666	0.669	0.674	0.682	0.689	0.696	0.703	5.48
BOL		0.641	0.647	0.654	0.658	0.662	0.667	0.671	0.677	0.683	0.689	0.695	4.96
GUY		0.624	0.63	0.629	0.634	0.636	0.637	0.64	0.644	0.648	0.651	0.655	3.00
NIC		0.619	0.623	0.625	0.628	0.631	0.636	0.643	0.651	0.66	0.668	0.676	7.13
GTM		0.611	0.617	0.624	0.626	0.627	0.632	0.639	0.647	0.655	0.663	0.671	7.05
HND		0.61	0.612	0.607	0.604	0.606	0.61	0.616	0.622	0.629	0.635	0.641	5.84
HTI		0.471	0.475	0.479	0.481	0.483	0.487	0.492	0.497	0.502	0.507	0.513	6.13
Latin America	μ	0.699	0.703	0.707	0.71	0.712	0.716	0.721	0.727	0.734	0.74	0.747	4.99
	±	0.081	0.081	0.083	0.083	0.083	0.083	0.083	0.084	0.084	0.085	0.086	3.90
Global	μ	0.679	0.683	0.687	0.689	0.691	0.695	0.7	0.705	0.711	0.717	0.723	4.63
	±	0.157	0.156	0.156	0.155	0.155	0.154	0.155	0.155	0.155	0.155	0.155	0.2

* Three-letter international country codes.

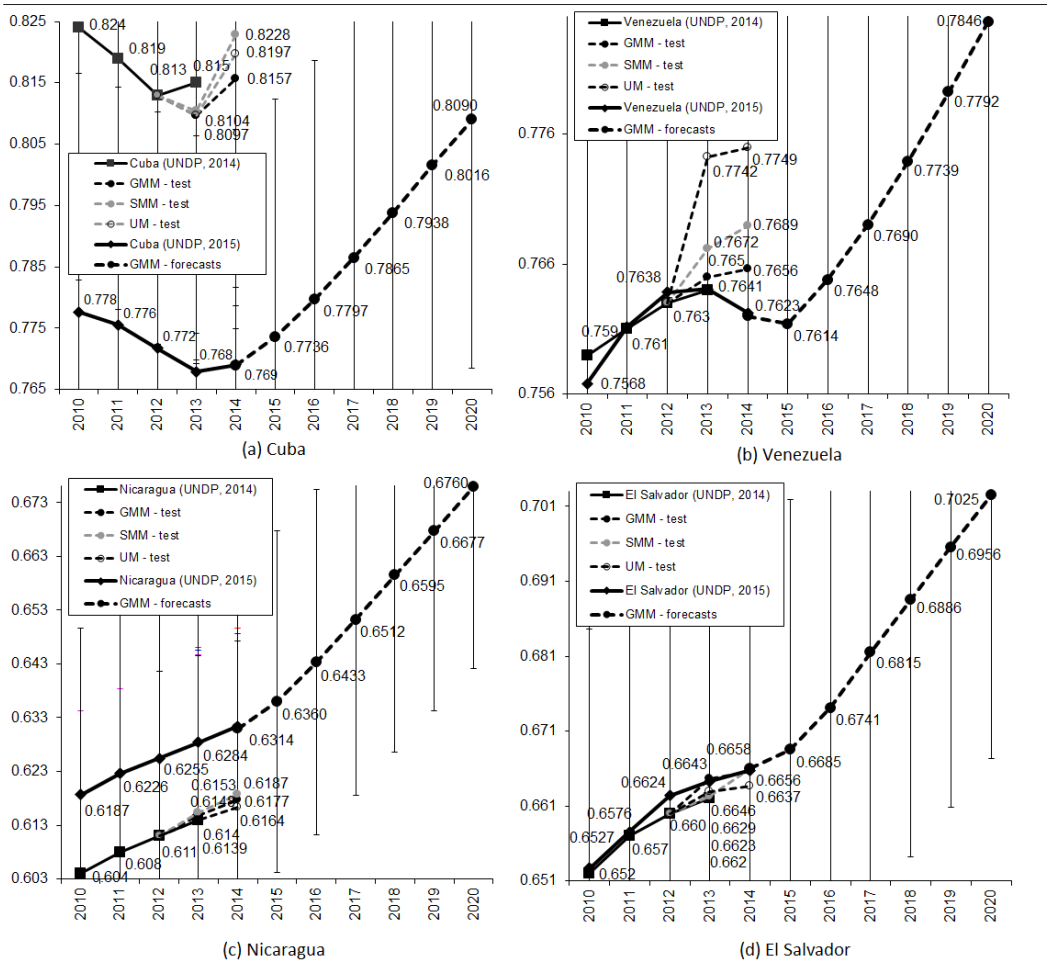

Figure 2. HDI forecasts and trends for Cuba (a), Venezuela (b), Nicaragua (c), and El Salvador (d).

Chile, currently classified[5] as developed countries with a very high level of human development (HDI > 0.799). The same may occur for Panama in 2017, Cuba in 2019 and Costa Rica in 2020. The developing countries of Paraguay and El Salvador may move from the medium human development class (0.599 < HDI < 0.7) to the high human development class (0.699 < HDI < 0.8) in 2019 and 2020, respectively.

The mean HDI of the Latin American countries (4.99 ± 3.90%) forecasted for the period from 2015 to 2020 shows an expected growth above the world average (4.63 ± 0.20%), maintaining the same trend[5] that shows Latin America and the Caribbean with the highest HDI, classified as high, and with indices higher than those in the regions of Europe, Asia, Pacific, the Arab States and sub-Saharan Africa.

Nicaragua (7.13%) and Guatemala (7.02%) show the highest index growth for the same period, while Belize (1.86%) shows the lowest growth. Haiti (6.13%), despite a growth trend above the world average, remains as the only Latin American country classified[5] as underdeveloped (HDI < 0.55). The other countries tend to remain in the same human development class, despite index growth.

Brazil, which is currently experiencing an economic crisis[40], will not see significant changes in its HDI despite significant advances being expected in LE and education because with the new calculation method, these advances tend to be mitigated by low income as a function of the GDP deficit. The geometric mean employed in the calculation of the index reduces the level of substitutability between dimensions, as low per-

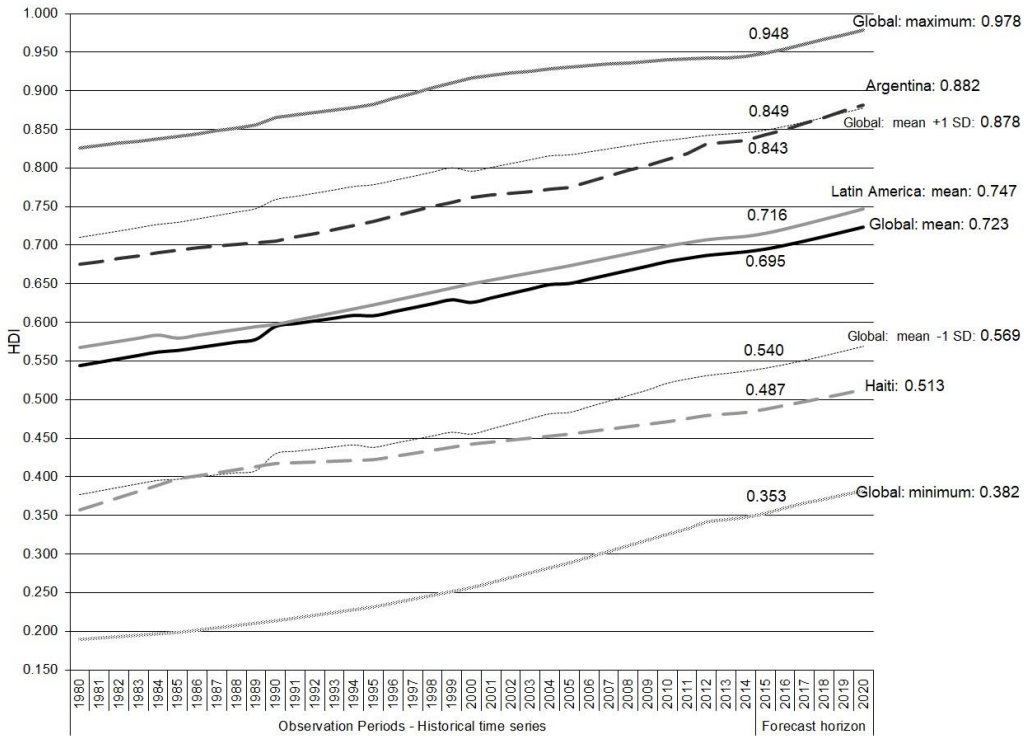

Figure 3. Historical time series (1980-2014) and forecast horizon (2015-2020) of the global HDI and HDI for Latin America and Latin American countries with the highest (Argentina) and lowest (Haiti) index.

formance in a given dimension can no longer be compensated for by better performance in another dimension[5]. Despite the criticisms[41] of the new HDI calculation method, it is observed that it favours countries with less inequality among its components[4], as the geometric mean tends to become increasingly smaller than the arithmetic mean as the variance between the components increases.

Life expectancy forecasts

It was possible to compare forecasts with recent studies from other international agencies[42,43] that had already reported the 2015 LE of its affiliated countries. However, their time series differ from the data source[3,5,37] used in the training of the models developed herein, which limits this study until new LE values or other studies are published that allow comparisons.

This indicator may also be inconsistent because many deaths are not recorded correctly[16].

When analysing the global historical LE series before forecasts for Latin America through the UNDP database[37], it is observed that in the last 34 years, the global average was 67.84 ± 2.89 years. The lowest LE recorded in the period was for Cambodia in 1980, with a mean of 27.5 years, while the LE of the world population in that period was 61.62 ± 10.5 years. In 1995, Rwanda had the lowest LE (31.50 years), well below the global average at the time of 65.44 ± 10.18 years. In the last report[5], Hong Kong - China recorded the highest LE (84 years in 2014) and Swaziland the lowest LE (49 years), while the global LE average was 71.03 ± 8.37 years. In Latin America, Haiti has always had the lowest LE, which was 62.8 years in 2014 and estimated, according to the forecasts herein, to reach 65.89 years by 2020, above the global average -1SD (65.06 years) forecasted. Other countries, such as Belize (70 years), Bolivia (68.3 years) and Guyana (66.4years), also feature LEs below the global average, while most, namely, 81.82% of Latin American countries,

Table 4. Latest observation points of historical time series of life expectancy in Latin American countries and forecasts for 2015-2020.

Country*		Latest Observation Points					Forecast Horizon						Variation %
		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2015-2020
CHL		80.4	80.7	81.1	81.4	81.7	82.14	82.63	83.19	83.81	84.50	85.25	4.35
CRI		78.8	78.9	79.1	79.2	79.4	79.58	79.82	80.09	80.40	80.75	81.13	2.18
CUB		79	79.1	79.2	79.3	79.4	79.53	79.64	79.78	79.95	80.17	80.41	1.27
PAN		76.8	77	77.2	77.4	77.6	77.86	78.12	78.43	78.76	79.14	79.56	2.52
URY		76.6	76.7	76.9	77	77.2	77.37	77.59	77.84	78.13	78.47	78.83	2.11
MEX		76.1	76.2	76.4	76.6	76.8	77.04	77.33	77.70	78.13	78.63	79.18	3.10
ARG		75.6	75.8	75.9	76.1	76.3	76.54	76.79	77.09	77.43	77.80	78.20	2.49
ECU		75	75.2	75.4	75.7	75.9	76.25	76.65	77.15	77.72	78.35	79.06	4.16
NIC		73.7	74	74.3	74.6	74.9	75.36	75.90	76.56	77.30	78.13	79.05	5.55
PER		73.7	73.9	74.1	74.3	74.6	74.93	75.36	75.89	76.51	77.22	77.99	4.55
BRA		73.3	73.6	73.9	74.2	74.5	74.91	75.35	75.86	76.43	77.08	77.79	4.41
VEN		73.6	73.7	73.9	74	74.2	74.38	74.61	74.85	75.14	75.46	75.81	2.17
COL		73.3	73.5	73.7	73.9	74	74.22	74.44	74.71	75.00	75.34	75.72	2.33
DOM		72.7	72.9	73.1	73.3	73.5	73.79	74.11	74.51	74.96	75.47	76.05	3.47
HND		72.4	72.6	72.8	72.9	73.1	73.37	73.69	74.09	74.58	75.16	75.81	3.71
SLV		71.9	72.2	72.5	72.8	73	73.41	73.94	74.59	75.30	76.11	77.03	5.52
PRY		72.3	72.5	72.6	72.8	72.9	73.09	73.25	73.47	73.69	73.97	74.27	1.88
GTM		70.9	71.1	71.4	71.6	71.8	72.14	72.56	73.07	73.66	74.34	75.11	4.61
BLZ		69.7	69.7	69.8	69.9	70	70.07	70.14	70.22	70.28	70.32	70.34	0.48
BOL		66.4	66.9	67.5	67.9	68.3	68.87	69.44	70.09	70.80	71.63	72.55	6.23
GUY		66	66.1	66.2	66.3	66.4	66.54	66.72	66.91	67.12	67.34	67.57	1.77
HTI		61.3	61.7	62.1	62.4	62.8	63.23	63.66	64.13	64.65	65.24	65.89	4.91
Latin America	μ	73.2	73.4	73.6	73.8	74	74.3	74.6	75	75.4	75.9	76.5	3.33
	±	4.47	4.43	4.39	4.38	4.36	4.34	4.33	4.33	4.35	4.37	4.42	2.05
Global	μ	69.9	70.2	70.5	70.8	71	71.4	71.7	72	72.4	72.8	73.3	3.18
	±	8.89	8.73	8.59	8.47	8.37	8.27	8.22	8.19	8.19	8.21	8.24	2.20

* International country code.

have an LE above the global average. The mean LE of Latin America[37,42,43] has historically always been higher than the global average.

The forecasts of this study estimate that in the next six years, the mean LE in Latin America will increase from 74 to 76.5 ± 4.42 years, while the global estimate is 73.29 ± 8.24 years, increasing to 74.3 ± 4.34 by 2015, as already confirmed in another study[42].

Currently[37], Chile (81.7 years) has the highest LE in Latin America, with a mean higher than that of other developed countries (79.9 ± 2.81), and is expected to reach an LE of 85.25 years by 2020, which is higher than that predicted in this study for the developed countries (81.61 ± 3.12 years).

Although LEs are increasing, Kanso et al.[14] notes that LE at age 60 would increase by 20% if preventable deaths did not occur and that male mortality was higher for almost all avoidable causes of death analysed, which may be related to men's greater exposure to risk factors and lower use of health services. On the other hand, studies[19] show a notable disadvantage of females regarding a healthy LE.

LE data, especially when broken down by sex and region, both with or without a health component, can be used in public policies as a reference for determining health plans and social security contributions[44] as well as provisions for pension payments[45,46], as justifications for social security reforms[47], in planning the future of

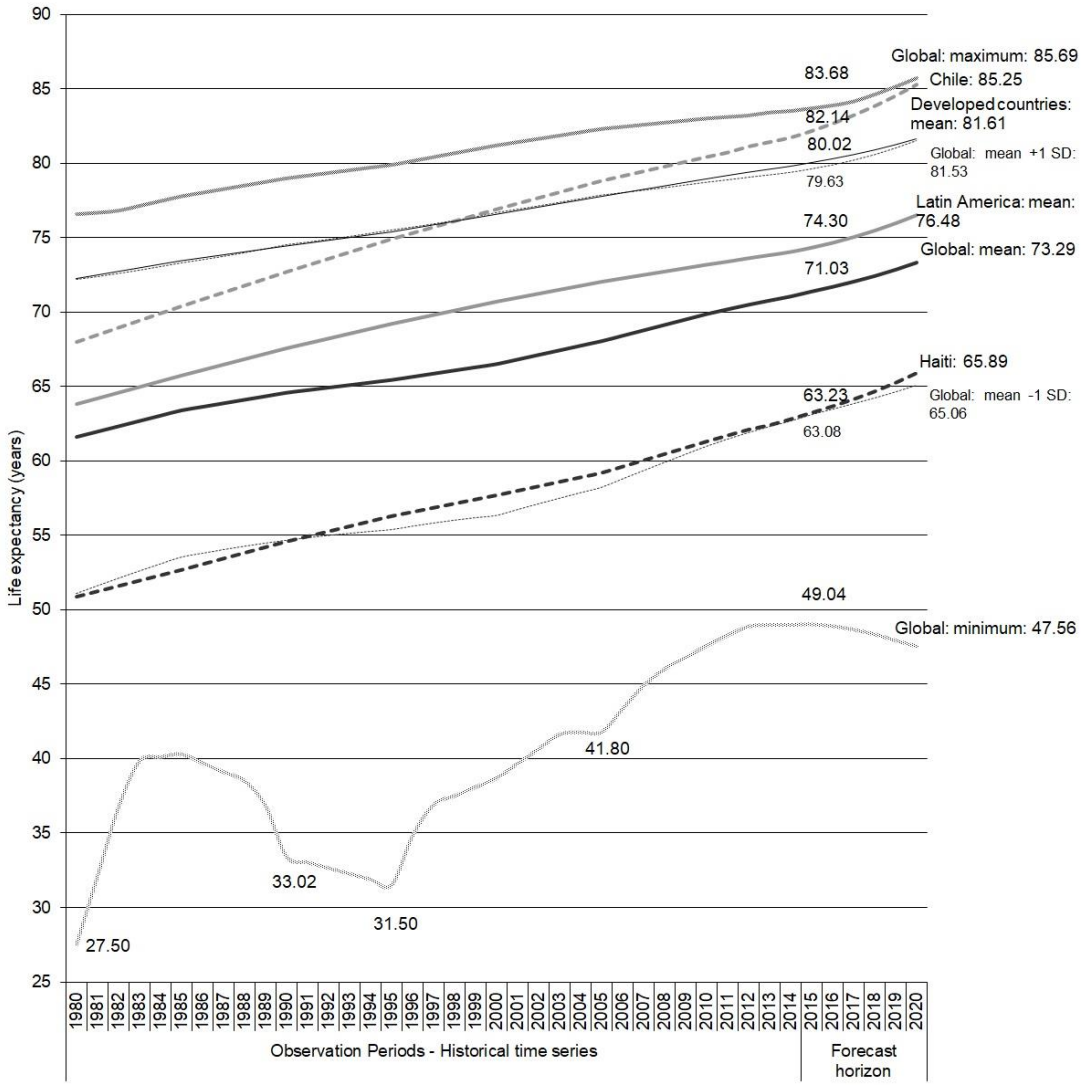

Figure 4. Historical time series (1980-2014) and forecast horizon (2015-2020) of the global LE and LE for Latin America and Latin American countries with the highest (Chile) and lowest (Haiti) LE.

health care[15] and in improving the quality of life of the elderly[48-50], as well as to predict any increase in age-related diseases[51].

Forecasting method, models, and algorithm

An extensive portion of the literature suggests that combined forecasts can improve individual ones[30]. This was visible in the multivariate models (MMs), which present better results than the UMs. In the MMs, the algorithm learned from the historical behaviour of the time series of all or groups of countries, while in the UMs, the learning was limited to the time series of the target country.

GMMs performed better than the other models. However, this relative advantage of the multivariate predictor may be different for each country. Other studies[52] also highlight the advantages of MMs, especially in cases of strong relationships between the time series, as observed in the present study.

The analysis of variance tests show no significant differences between model predictions and the trends reported by the UNDP[5] for the 2013 and 2014 HDI. However, the GMMs presented

the best cumulative quality measures during the entire training and forecasting period, with the highest DAC and the lowest errors relative to the other models.

The efficiency of GMMs can be implicitly explained by the interdependencies and vulnerability of countries, as stated in other studies[3].

Regarding the method of evaluation of the models, it is observed that the MAE and DAC quality measures are sufficient to qualify the HDI or LE forecast, precluding the need for an analysis of variance since, although there are no significant differences between model predictions and the actual values, the MAE allowed the identification of the best models, confirming studies[53] that discuss the use of specific forecast quality measures.

The SMOReg algorithm presented the best quality measures during the prediction tests relative to the other function-based learning algorithms (Table 1), confirming previous studies[54] and reaffirming the advantages of using DM techniques (Figure 1) compared to other more popular forecasting techniques, such as ARIMA, analysed in previous studies[55].

The major difficulty of this forecasting method is the operational cost. The DM pre- and post-processing stages consumed approximately 80% of the operational cost, as previous studies suggest[56]. Lack of access to consistent data was another problem, frequent in large databases[57], as updates to previously published observational data limited the study, decreasing its predictive ability.

Conclusion

Models developed from multivariate time series, although more complex, presented better accuracy than the models developed from univariate series.

The multivariate time series allow greater learning of the algorithms with the increase of different univariate historical experiences.

Data mining techniques provided better forecasts than the most popular technique, ARIMA.

The HDI is a robust index with great predictability and vulnerability, used in epidemiological research, mainly as a demographic delimiter or a comparative parameter.

The mean HDI and LE growth in Latin American countries is expected to remain higher than the global average.

The contradictions between the forecasted and actual values of the index or its components, if compared, may in the future trigger discussions and help in decision-making to support public policies regarding health planning and management and explain the scenarios observed in countries and the world.

Collaborations

CB Santos - contributed substantially to the study design and planning, contributed to the analysis and interpretation of data, and approved the final version of the manuscript. LA Pilatti, B Pedroso, DR Carvalho and AM Guimarães – contributed substantially to the study design and planning, contributed to the critical review of content, and approved the final version of the manuscript.

References

1. United Nations Development Programme (UNDP). *Human Development Report (HDR) 1990: Concept and Measurement of human development*. New York: UNDP; 1990.
2. Alkire S. Human development: Definitions, critiques, and related concepts. *UNDP-HDRO Occasional Papers* 2010.
3. United Nations Development Programme (UNDP). *Human Development Report (HDR) 2014. Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience*. New York: UNDP; 2014.
4. Kovacevic M. Review of HDI critiques and potential improvements, UNDP. *Human Development Reports* 2010; 33.
5. United Nations Development Programme (UNDP). *Human Development Report (HDR) 2015. Rethinking Work for Human Development*. New York: UNDP; 2015.
6. Percio J, Medina NH, Luna EA. Visual Impairment and Human Development in Brazil. *Int J Epidemiol* 2015; 44(Supl. 1):i157.
7. Sadovsky ADI, Poton WL, Reis-Santos B, Barcelos MRB, Silva ICM. Índice de Desenvolvimento Humano e prevenção secundária de câncer de mama e colo do útero: um estudo ecológico. *Cad Saude Publica* 2015; 31(7):1539-1550.
8. Tavares LF, Castro IRR, Levy RB, Cardoso LO, Claro RM. Dietary patterns of Brazilian adolescents: results of the Brazilian National School-Based Health Survey (PeNSE). *Cad Saude Publica* 2014; 30(12):2679-2690.
9. Szuster DAC, Caiaffa WT, Andrade EIG, Acurcio FA, Cherchiglia ML. Sobrevida de pacientes em diálise no SUS no Brasil. *Cad Saude Publica* 2012; 28(3):415-424.
10. Castro JMd, Rodrigues-Júnior AL. A influência da mortalidade por causas externas no desenvolvimento humano na Faixa de Fronteira brasileira. *Cad Saude Publica* 2012; 28(1):195-200.
11. Kariminia A, Chokephaibulkit K, Pang J, Lumbiganon P, Hansudewechakul R, Amin J, Kumarasamy N, Puthanakit T, Kurniati N, Nik Yusoff NK, Saphonn V, Fong SM, Razali K, Nallusamy R, Sohn AH, Sirisanthana V. Cohort Profile: The TREAT Asia Pediatric HIV Observational Database. *Int J Epidemiol* 2011; 40(1):15-24.
12. Martinez EZ, Roza DL, Caccia-Bava MCGG, Achcar JA, Dal-Fabbro AL. Gravidez na adolescência e características socioeconômicas dos municípios do Estado de São Paulo, Brasil: análise espacial. *Cad Saude Publica* 2011; 27(5):855-867.
13. González-Zapata LI, Estrada-Restrepo A, Álvarez-Castaño LS, Álvarez-Dardet C, Serra-Majem L. Exceso de peso, aspectos económicos, políticos y sociales en el mundo: un análisis ecológico. *Cad Saude Publica* 2011; 27(9):1746-1756.
14. Kanso S, Romero DE, Leite IC, Marques A. A evitabilidade de óbitos entre idosos em São Paulo, Brasil: análise das principais causas de morte. *Cad Saude Publica* 2013; 29(4):735-748.
15. Mendes ACG, Sá DA, Miranda GMD, Lyra TM, Tavares RAW. Assistência pública de saúde no contexto da transição demográfica brasileira: exigências atuais e futuras. *Cad Saude Publica* 2012; 28(5):955-964.
16. Chiavegatto Filho ADP, Laurenti R. Decomposição da diferença da expectativa de vida de Minas Gerais em relação ao Rio de Janeiro e São Paulo, Brasil. *Cad Saude Publica* 2013; 29(6):1131-1140.
17. Cervantes CAD, Botero MA. Average years of life lost due to breast and cervical cancer and the association with the marginalization index in Mexico in 2000 and 2010. *Cad Saude Publica* 2014; 30(5):1093-1102.
18. Campolina AG, Adami F, Santos JLF, Lebrão ML. A transição de saúde e as mudanças na expectativa de vida saudável da população idosa: possíveis impactos da prevenção de doenças crônicas. *Cad Saude Publica* 2013; 29(6):1217-1229.
19. Camargos MCS, Gonzaga MR. Viver mais e melhor? Estimativas de expectativa de vida saudável para a população brasileira. *Cad Saude Publica* 2015; 31(7):1460-1472.
20. Stringhini S, Polidoro S, Sacerdote C, Kelly RS, van Veldhoven K, Agnoli C, Grioni S, Tumino R, Giurdanella MC, Panico S, Mattiello A, Palli D, Masala G, Gallo V, Castagné R, Paccaud F, Campanella G, Chadeau-Hyam M, Vineis P. Life-course socioeconomic status and DNA methylation of genes regulating inflammation. *Int J Epidemiol* 2015; 44(4):1320-1330.
21. Li L, Hardy R, Kuh D, Power C. Life-course body mass index trajectories and blood pressure in mid life in two British birth cohorts: stronger associations in the later-born generation. *Int J Epidemiol* 2015; 44(3):1018-1026.
22. Lacey RE, Sacker A, Kumari M, Worts D, McDonough P, Booker C, McMunn A. Work-family life courses and markers of stress and inflammation in mid-life: evidence from the National Child Development Study. *Int J Epidemiol* 2015; 45(4):1247-1259.
23. Hendi AS. Trends in U.S. life expectancy gradients: the role of changing educational composition. *Int J Epidemiol* 2015; 44(3):946-955.
24. Morton SM, De Stavola BL, Leon DA. Intergenerational determinants of offspring size at birth: a life course and graphical analysis using the Aberdeen Children of the 1950s Study (ACONF). *Int J Epidemiol* 2014; 43(3):749-759.
25. Anstey KJ, Kingston A, Kiely KM, Luszcz MA, Mitchell P, Jagger C. The influence of smoking, sedentary lifestyle and obesity on cognitive impairment-free life expectancy. *Int J Epidemiol* 2014; 43(6):1874-1883.
26. Brunekreef B, Von Mutius E, Wong GK, Odhiambo JA, Clayton TO, Group tIPTS. Early life exposure to farm animals and symptoms of asthma, rhinoconjunctivitis and eczema: an ISAAC Phase Three Study. *Int J Epidemiol* 2012; 41(3):753-761.
27. Mackenbach JP, Looman CW. Life expectancy and national income in Europe, 1900-2008: an update of Preston's analysis. *Int J Epidemiol* 2013; 42(4):1100-1110.
28. Mangalova E, Agafonov E. Wind power forecasting using the k-nearest neighbors algorithm. *Int J Forecasting* 2014; 30(2):402-406.
29. Silva L. A feature engineering approach to wind power forecasting: GEFCom 2012. *Int J Epidemiol* 2014; 30(2):395-401.

30. Rodrigues BD, Stevenson MJ. Takeover prediction using forecast combinations. *Int J Forecasting* 2013; 29(4):628-641.
31. Correa FE, Gama J, Pizzigatti Correa PL, Alves LRA. Data mining frequent temporal events in agrieconomic time series. *IEEE Lat Am T* 2015; 13(7):2329-2334.
32. Sousa WRN, Couto MS, Castro AF, Silva MPS. Evaluation of desertification processes in ouricuri-pe through trend estimates of times series. *IEEE Lat Am T* 2013; 11(1):602-606.
33. Xie Y, Schreier G, Hoy M, Liu Y, Neubauer S, Chang DCW, Redmond SJ, Lovell NH. Analyzing health insurance claims on different timescales to predict days in hospital. *J Biomed Inform* 2016; 60:187-196.
34. Winters-Miner LA, Bolding PS, Hilbe JM, Goldstein M, Hill T, Nisbet R, Walton N, Miner GD. Biomedical Informatics. In: Winters-Miner LA, Bolding PS, Hilbe JM, Goldstein M, Hill T, Nisbet R, Walton N, Miner GD. *Practical Predictive Analytics and Decisioning Systems for Medicine*. Cambridge: Academic Press; 2015. p. 42-59.
35. Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI magazine* 1996; 17(3):37.
36. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 2009; 11(1):10-18.
37. United Nations Development Programme (UNDATA). *Human Development Index trends, 1980–2013*. New York: UNDATA; 2014.
38. Hall MA. *Correlation-based feature selection for machine learning* [thesis]: The University of Waikato; 1999.
39. Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK. Improvements to the SMO algorithm for SVM regression. *IEEE T Neur Net Lear* 2000; 11(5):1188-1193.
40. Watts J. Brazil's health system woes worsen in economic crisis. *Lancet* 2016; 387(10028):1603-1604.
41. Ravallion M. Troubling tradeoffs in the human development index. *J Dev Econ* 2012; 99(2):201-209.
42. Central Intelligence Agency (CIA). *The World Factbook*. Langley: CIA; 2016.
43. World Bank. *World Development Indicators*. Washington: The World Bank; 2016.
44. Inoue JT, Rodrigues CG, Afonso LE. Tábua de mortalidade e expectativa de vida saudável: uma aplicação à população beneficiária de planos de saúde privados no Brasil em 2008. In: *Anais do 12º Congresso USP de Controladoria e Contabilidade*; 2012; São Paulo. p. 1-15.
45. Brasil. Lei nº 13.183, de 4 de novembro de 2015. Altera as Leis nºs 8.212, de 24 de julho de 1991, e 8.213, de 24 de julho de 1991, para tratar da associação do segurado especial em cooperativa de crédito rural e, ainda essa última, para atualizar o rol de dependentes, estabelecer regra de não incidência do fator previdenciário, regras de pensão por morte e de empréstimo consignado, a Lei nº 10.779, de 25 de novembro de 2003, para assegurar pagamento do seguro-defeso para familiar que exerça atividade de apoio à pesca, a Lei nº 12.618, de 30 de abril de 2012, para estabelecer regra de inscrição no regime de previdência complementar dos servidores públicos federais titulares de cargo efetivo, a Lei nº 10.820, de 17 de dezembro de 2003, para dispor sobre o pagamento de empréstimos realizados por participantes e assistidos com entidades fechadas e abertas de previdência complementar e a Lei nº 7.998, de 11 de janeiro de 1990; e dá outras providências. *Diário Oficial da União* 2015; 5 nov.
46. Lu B, He W, Piggott J. Should China introduce a social pension? *The Journal of the Economics of Ageing* 2014; 4:76-87.
47. Rocha FRF. A previdência social no Brasil: uma política em reestruturação. *Temporalis* 2016; 2(30):453-473.
48. Rosa VD. Atividade física e a qualidade de vida de mulheres idosas. *FACES* 2016.
49. Vecchia RD, Ruiz T, Bocchi SCM, Corrente JE. Qualidade de vida na terceira idade: um conceito subjetivo. *Rev Bras Epidemiol* 2005; 8(3):246-252.
50. Minayo MCS, Hartz ZMA, Buss PM. Qualidade de vida e saúde: um debate necessário. *Cien Saude Colet* 2000; 5(1):7-18.
51. Salgado Filho N, Brito DJA. Doença renal crônica: a grande epidemia deste milênio. *J Bras Nefrol* 2006; 28(2):1-5.
52. Peña D, Sánchez I. Measuring the advantages of multivariate vs. univariate forecasts. *J Time Ser Anal* 2007; 28(6):886-909.
53. Armstrong JS. Evaluating forecasting methods. In: Armstrong JS, editor. *Principles of forecasting*: Springer; 2001. p. 443-472.
54. Li C, Jiang L. Using locally weighted learning to improve SMOreg for regression. *Pacific Rim International Conference on Artificial Intelligence*; 2006; Berlin Heidelberg: Springer; 2006. p. 375-384.
55. Hong T, Pinson P, Fan S. Global energy forecasting competition 2012. *Int J Forecasting* 2014; 30(2):357-363.
56. Mannila H. Data mining: machine learning, statistics, and databases. *Scientific and Statistical Database Management, International Conference on*; 1996: IEEE Computer Society; 1996. p. 2-2.
57. Witten IH, Frank E. *Practical machine learning tools and techniques*. 2nd ed. San Francisco: Morgan Kaufmann; 2005.

Article submitted 05/25/2016
Approved 10/21/2016
Final version submitted 10/23/2016