PREVISÃO DO ÍNDICE DE DESENVOLVIMENTO HUMANO DE 2013 E 2014 POR MEIO DE TÉCNICAS DE MINERAÇÃO DE DADOS EM SÉRIES TEMPORAIS UNIVARIADAS E MULTIVARIADAS

Celso Bilynkievycz dos Santos, Bruno Pedroso, Alaine Margarete Guimarães, Luiz Alberto Pilatti e João Luiz Kovaleski

RESUMO

O Índice de Desenvolvimento Humano (IDH) é um indicador adotado pela Organização Mundial da Saúde para avaliar a qualidade de vida de uma determinada região. Sua previsão pode auxiliar no planejamento e tomada de decisões para orientação e defesa de políticas para melhorar o seu desenvolvimento. Este estudo fez a previsão do IDH de 2013 e 2014 a partir de técnicas de mineração de dados de previsão em séries temporais, perfazendo todas as etapas do processo de descoberta de conhecimento em bases de dados. No estudo foi avaliada a capacidade preditiva de 376 modelos, dois genéricos e 374 específicos por país. Para o desenvolvimento dos modelos foi utilizado o algoritmo SMOReg, executado em uma aplicação de interface de programação Forecast *do ambiente WEKA. O modelo genérico foi treinado e testado com séries temporais multivariadas, correspondentes aos registros de IDH de 187 países, enquanto que os modelos específicos formam desenvolvidos a partir de séries temporais univariadas, correspondentes ao comportamento histórico individual do índice em cada país. As variáveis temporais utilizadas corresponderam aos períodos históricos e intermitentes de 1980 a 2013 publicados no relatório do Programa das Nações Unidas para o Desenvolvimento em 24/07/2014. Na análise empírica verificou-se que os modelos multivariados apresentaram as melhores medidas de qualidade nas previsões. As previsões do IDH 2013 foram eficientes, não apresentando diferenças significativas dos valores publicados, enquanto as previsões do IDH 2014 dependem de comparação com os valores divulgados posteriormente à finalização do presente trabalho.*

Introdução

Profissionais de desenvolvimento humano, ao avaliarem diferentes aspectos e consequências dos países em crescimento, perceberam que o desenvolvimento não deve ser visto apenas como mera expansão do seu crescimento econômico (referencias). Desde os anos 1970, essa percepção cresceu, observando que os esforços investidos na industrialização e crescimento econômico não conduziram a uma redução considerável da pobreza e das desigualdades nos países em desenvolvimento, pois, somente o desenvolvimento econômico não conseguiria atender as necessidades básicas das camadas mais pobres da população, como água, eletricidade, saúde e educação. Observou-se também que em algumas áreas os indicadores sociais pioraram, enquanto o produto interno bruto (PIB) global apresentava taxas de crescimento significativas. Estas críticas levaram à necessidade da criação de um indicador de desenvolvimento humano.

Em 1990, um grupo de economistas formado por Mahbub ul Haq, Amartya Sen, Paul Streeten e Keith Griffin criou o Índice de Desenvolvimento Humano (IDH), que vem sendo utilizado pelo Programa das Nações Unidas para o Desenvolvimento (UNDP, do inglês *United Nations Development Programme*) para avaliar o desenvolvimento humano dos países filiados (UNDP, 1990).

A previsibilidade do IDH pode auxiliar em tomadas de decisões governamentais, e, caso as expectativas não correspondessem aos valores reais, poderá apoiar ou não medidas políticas ou econômicas.

A literatura oferece uma variedade de técnicas de previsão, entre elas destacam-se as previsões a partir de técnicas de mineração de dados (MD) aplicadas em séries temporais.

Diferentes estudos contemporâneos de previsão utilizando técnicas de MD foram desenvolvidos em diferentes áreas, entre elas, energia eólica (Mangalova e Agafonov, 2014; Silva, 2014), mercado financeiro (Rodrigues e Stevenson, 2013) e engenharia (Palit e Popovic, 2006; Alonso, Cruz e Barceló, 2009).

KEYWORDS / Anthocyanins / Indicadores de Desenvolvimento / Mineração de Dados / Previsões /

Recebido: 17/04/2015. Modificado: 16/01/2019. Aceito: 17/01/2019.

Celso Bilynkievycz dos Santos. Mestrado e Doutorado em Engenharia da Produção, Universidade Tecnológica Federal do Paraná (UTFPR), Brasil. Professor, Universidade Estadual de Ponta Grossa (UEPG), Brasil. Endereço: Setor de Ciências Biológicas e da Saúde, UEPG. Av. General Carlos Cavalcanti, 4748. Uvaranas 84030900. Ponta Grossa, PR, Brasil. e-mail: bilynkievycz@uepg.br

Bruno Pedroso. Mestrado em Engenharia da Produção, UTFPR, Brasil. Doutorado em Educação Física, Universidade Estadual de Campinas (Unicamp), Brasil. Professor, UEPG, Brasil. e-mail: prof.brunopedroso@gmail.com

Alaine Margarete Guimarães. Mestrada em Informática, Universidade Federal do Paraná, Brasil. Doutorada em Agronomía, Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Brasil. Professora, UEPG, Brasil. e-mail: alainemg@uepg.br

Luiz Alberto Pilatti. Mestrado em Educação, Universidade Metodista de Piracicaba, Brasil. Doutorado em Educação Física, Unicamp, Brasil. Professor, UTFPR, Brasil. Bolsista CNPq, Brasil. e-mail: lapilatti@utfpr.edu.br.

João Luiz Kovaleski. Mestrado em Ciências, UTFPR, Brasil. Doutorado em Instrumentação Industrial, Université Joseph Fourier, França. Professor, UTFPR, Brasil. Bolsista CNPq, Brasil. e-mail: kovaleski@utfpr.edu.br.

PREVISIÓN DEL ÍNDICE DE DESARROLLO HUMANO DE 2013 Y 2014 POR MEDIO DE TÉCNICAS DE MINERÍA DE DATOS EN SERIES TEMPORALES UNIVARIADAS Y MULTIVARIADAS

Celso Bilynkievycz dos Santos, Bruno Pedroso, Alaine Margarete Guimarães, Luiz Alberto Pilatti y João Luiz Kovaleski

RESUMEN

El Índice de Desarrollo Humano (IDH) es un indicador adoptado por la Organización Mundial de la Salud para evaluar la calidad de vida de una región determinada. Su previsión puede ayudar en la planificación y toma de decisiones para orientación y defensa de políticas para mejorar su desarrollo. Este estudio hizo la previsión del IDH de 2013 y 2014 a partir de técnicas de minería de datos de previsión en series temporales, incluyendo todas las etapas del proceso de descubrimiento de conocimiento en bases de datos. Se evaluó la capacidad predictiva de 376 modelos, dos genéricos y 374 específicos, por país. Para el desarrollo de los modelos se utilizó el algoritmo SMOReg, ejecutado en una aplicación de interfaz de programación Forecast del ambiente WEKA. El modelo genérico fue entrenado y probado con series temporales multivariadas, correspondientes a los registros de IDH de 187 países, mientras que los modelos específicos se desarrollan a partir de series temporales univariadas, correspondientes al comportamiento histórico individual del índice en cada país. Las variables temporales utilizadas corresponden a los períodos históricos e intermitentes de 1980 a 2013 publicados en el informe del Programa de las Naciones Unidas para el Desarrollo el 24/07/2014. En el análisis empírico se verificó que los modelos multivariados presentaron las mejores medidas de calidad en las previsiones. Las previsiones del IDH de 2013 fueron eficientes, no presentando diferencias significativas con los valores publicados, mientras que las previsiones del IDH 2014 dependen de comparación con los valores divulgados después de la finalización del presente trabajo.

FORECASTING THE HUMAN DEVELOPMENT INDEX OF 2013 AND 2014 BY MEANS OF DATA MINING TECHNIQUES IN UNIVARIATE AND MULTIVARIATE TEMPORARY SERIES

Celso Bilynkievycz dos Santos, Bruno Pedroso, Alaine Margarete Guimarães, Luiz Alberto Pilatti and João Luiz Kovaleski

SUMMARY

The Human Development Index (HDI) is an indicator adopted by the World Health Organization to assess the quality of life of a given region. Its prediction can aid in planning and decision-making for policy guidance and advocacy to improve its development. This study predicted the HDI of 2013 and 2014 from forecasting data mining techniques in time series, completing all stages of the knowledge discovery process in databases. In the study, the predictive capacity of 376 models, two generic and 374 country specific, were evaluated. For the development of the models we used the SMOReg algorithm, executed in a Forecast programming interface application of the WEKA environment. The generic model was trained and tested with multivariate time series corresponding to the HDI records of 187 countries, while the specific models were developed from univariate time series corresponding to the individual historical behavior of the index in each country. The time variables used corresponded to historical and intermittent periods from 1980 to 2013 published in the report of the United Nations Development Program on 07/24/2014. In the empirical analysis it was verified that the multivariate models presented the best quality measures in the predictions. The predictions of the HDI 2013 were efficient, with no significant differences to published figures, while the predictions of HDI 2014 depend on comparison with figures released after the completion of the present study.

A equipe de Silva (2014) desenvolveu um modelo preditivo de potência em usinas de energia eólica, utilizando os algoritmos de MD *Multiple linear regression*, GBM e K *Means*. Para o mesmo propósito, Mangalova e Agafonov (2014) combinaram métodos heurísticos e formais, utilizando o algoritmo CaRT (*Classification and regression tree*) na seleção de fatores.

No estudo de Rodrigues e Stevenson (2013) foram utilizados modelos de redes neurais artificiais (RNAs), regressão linear (RL) e combinações entre ambos para identificar na bolsa de valores, de forma precoce, empresas alvo de aquisições ou fusão. Seus resultados apontaram melhor desempenho individual dos modelos de RNAs em relação aos modelos de RL. Mas as combinações dos modelos superaram todos os resultados.

Palit e Popovic (2006) fizeram a análise das séries temporais apresentando modelos e aplicações de previsão em engenharia a partir de métodos matemáticos não convencionais da inteligência artificial (IA). Também, apresentaram taxonomia para séries temporais, de acordo com suas características.

Estudos de duas décadas atrás, como de Arinze (1994) já recomendavam algoritmos de MD para aquisição de conhecimento, reduzindo a necessidade de especialistas. No entanto, estudos posteriores fazem julgamentos desfavoráveis à utilização de técnicas de MD (Chatfield, 1995; Keogh e Kasetty, 2003; Armstrong, 2006). Keogh e Kasetty (2003) ainda criticam a falta de pesquisadores da área de MD para testar métodos alternativos.

A inexistência de estudos de MD, respeitando todas as etapas do processo de descoberta de conhecimento em base de dados (KDD, do inglês *knowledge discovery in databases*) (Fayyad *et al*., 1996) e envolvendo bases de dados reais de series temporais univariadas e multivariadas, com índices de baixa variabilidade temporal e baixa variância entre os períodos, mas com diferenças significativas entre eles, aponta uma lacuna existente na literatura, principalmente aplicando os algoritmos de aprendizagem baseada em funções por meio de uma interface de programação de aplicativos (API, do inglês *application programming interface*) recentemente disponibilizada para testes.

Diante desta abertura na literatura, este trabalho tem como objetivo realizar a previsão do IDH de 2013 e 2014 a partir de seus dados históricos, utilizando a técnica de MD *Forecast* em séries temporais univariadas e multivaridas. Os índices de IDH e as técnicas empregadas apresentam as características necessárias para contribuir no fechamento desta lacuna da literatura, como: i) Séries temporais univariadas correspondentes a cada país. ii) Séries temporais multivariadas correspondentes ao grupo de países filiados a Organização das Nações Unidas (ONU). iii) Variabilidade temporal baixa, de apenas 12 anos, correspondentes a um período intermitente entre 1980 a 2013. iv) Baixa variância média anual do índice por país (σ^2= 0,0005 ±0,00045). v) Diferença significativa do índice entre os tempos das séries temporais pareadas por países ($p<0,001$). vi) Utilização de algoritmos de aprendizagem baseada em funções. vii) Execução do algoritmo usando a API *Forecast* do WEKA, disponibilizada para comunidade científica no segundo semestre de 2013. viii) Realização da análise dos dados respeitando todas as etapas do KDD.

Materiais e Métodos

A partir dos dados dos IDH dos países filiados a ONU, foram desenvolvidas todas as etapas do processo de KDD (Fayyad *et al*., 1996) no ambiente de MD WEKA (Witten e Frank, 2005). Também, foram realizados testes estatísticos complementares de análise de variância, correlação e regressão em diferentes momentos do processo de KDD.

O hardware utilizado foi composto de um processador de 2.4GHz e 10GBytes de memória RAM dedicada ao processamento.

Para avaliar os resultados utilizou-se todas medidas de qualidade das previsões das séries temporais disponíveis na API do WEKA: erro absoluto médio (MAE, do inglês *mean absolute error*); erro quadrático médio (MSE, do inglês *mean squared error*); raiz quadrada do erro quadrático médio (RMSE, do inglês *root mean squared error*); erro percentual absoluto médio (MAPE, do inglês *mean absolute percentage error*); precisão direcional (DAC, do inglês *directional accuracy*); erro absoluto relativo (RAE, do inglês relative absolute error) e raiz quadrada do erro quadrático relativo (RRSE, do inglês *root relative squared error*).

Os parâmetros experimentais utilizados no processo de KDD estão apresentados na Tabela I, assim como a síntese das séries temporais utilizadas encontram-se na Tabela II, para possíveis replicações ou comparações entre pesquisas.

Previsão das tendências do IDH 2013 e 2014

O *Forecast* se diferencia dos demais métodos de classificação de MD por trabalhar com séries temporais, as quais consistem em conjuntos de observações de variáveis com dependência serial, ordenadas em função do tempo.

A previsão foi desenvolvida seguindo as etapas do processo de KDD (Fayyad *et al*., 1996), que são apresentadas na Figura 1 e estão divididas, segundo definições de Michalski e Kaufman (1998), em três macro-etapas: i) pré-processamento de MD: todas as subetapas que antecedem a de MD; ii) MD: etapa onde se aplicam os algoritmos mineradores; iii) pós-processamento de MD: todas as subetapas utilizadas para se consolidar o conhecimento.

Pré-processamento de MD

O pré-processamento iniciou-se com a obtenção dos dados

TABELA I
PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS KKD *

Procedimentos: Etapas do processo de KDD.
Fonte de Dados: Relatório da UNDP (2014) e site da UNDATA (2014) com séries temporais.
Softwares: WEKA, MS ACCESS, MS Excel, GraphPad InStat.
API: Forecast.
Problema de KDD: Classificação.
Tipo de Dados: Séries temporais univariadas e multivariadas.
Técnica de MD: Previsão.
Algoritmo: SMOReg - Selecionado entre algoritmos de aprendizagem baseada em funções.
Paradigma de aprendizagem dos algoritmos: Aprendizagem baseada em funções.
Avaliação dos resultados: Medidas de qualidade para *Forecast*.
Técnicas estatísticas: Kolmogorov-Smirnov, análise de variância, regressão linear, correlação.
Testes estatísticos paramétricos: correlação de Pearson, teste T pareado, ANOVA.
Testes estatísticos não-paramétricos: correlação de Spearman, Wilcoxon *matched-pairs* e Friedman.

* Do inglês, *knowledge discovery in databases*.

TABELA II
SÉRIES TEMPORAIS UTILIZADAS

Classe	Quant.	Nome	Unidades	Período	Previsão	Registros		Características predominantes	Descrição
						N	%		
Multivariada	01	IDH dos países	Índice	Anual (1980-2012) intermitente	IDH 2013	1982	88,32	Não estacionárias; Não sazonais; Linear	Índice de Desenvolvimento Humano dos 187 Países filiados a ONU
	01			Anual (1980-2013) intermitente	IDH 2014	2169	89,22		
Univariada	187	IDH por país		Anual (1980-2012) intermitente	IDH 2013	μ=11,60 ±2,41 (4-13)	μ=89,23 ±18,54 (30,77-100)	Não estacionárias; Não sazonais; Linear/Não Linear	Índice de Desenvolvimento Humano de cada país filiado a ONU
	187			Anual (1980-2013) intermitente	IDH 2013	μ=12,60 ±2,41 (5-14)	μ=90,00 ±18,54 (35,71-100)		

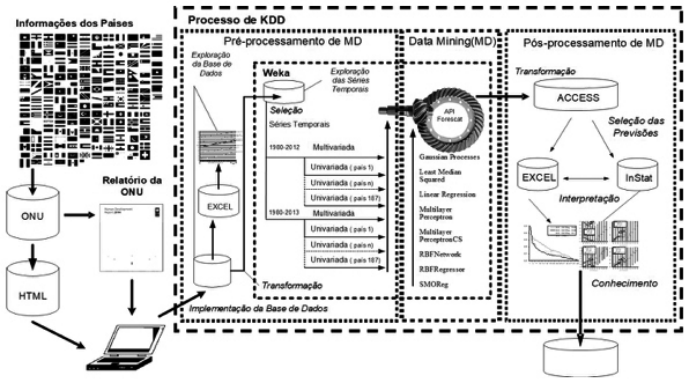

Figura 1. Etapas de desenvolvimento da pesquisa.

extraídos do relatório da ONU (UNDP, 2014) e de sua base de dados (UNDATA, 2014). A partir destas fontes foi desenvolvida uma base de dados específica com as séries temporais atualizadas em 24 de julho de 2014. Após a implementação desta base de dados, foi realizada a etapa de KDD de exploração da base de dados por meio da linguagem de consultas estruturadas (SQL, do inglês *structured query language*), resultando na estatística descritiva das séries temporais (Tabela III).

Observa-se a partir da Tabela III, que o IDH do países filiados, ao logo do período apresenta média crescente, e que o desvio padrão é decrescente. Os dados apresentam alta homogeneidade (baixa dispersão ou variabilidade). Constatou-se também, que 65,78% dos países apresentavam dados completos nas suas séries temporais. A ausência de dados é retroativa a 2010 (Figura 2). A ausência de dados foi tratada de forma não supervisionada, não tendo sido adotadas medidas supervisionadas.

Em seguida, os dados foram transformados no formato *attribute-relation file format* (arff) e continuou-se a exploração das séries temporais no ambiente WEKA.

A partir da mineração visual, da estatística descritiva, e da análise de correlação, as séries temporais foram caracterizadas, segundo as definições de Palit e Popovic (2006) em: i) Não estacionárias: apresentam comportamento de crescimento; (ii) Não sazonais: não apresentam padrões de comportamento em períodos regulares de tempo; e iii) Lineares: apresentam correlação com o tempo.

Observou-se também, que as séries multivariadas apresentavam alta correlação entre si, exceto com as séries temporais de cinco países que se apresentaram como *outliers* (Congo, República Democrática do Congo, Lesoto, Suazilândia e Zimbábue) e que também apresentaram características de não linearidade com o tempo.

A literatura apresenta diferentes técnicas de detecção de anomalias em séries temporais. Fox (1972) introduz técnicas estatisticamente rigorosas para tratar questões de elementos estranhos em séries temporais. Outros estudos apresentam diferentes tipos de elementos estranhos em séries (Chen e Liu, 1993a, b; Muirhead, 1986), onde são propostas técnicas para detectá-los e para obtenção de boas estimativas. Ainda, Abraham e Chuang (1989) propõem a utilização de redes bayesianas para o tratamento de elementos estranhos.

A presença de *outliers* afeta a regressão, porque o quadrado da distância mínima se acentua com a influência dos pontos mais distantes a partir da linha de regressão (Witten e Frank, 2005). A prática sugere eliminação desses *outliers* de forma supervisionada ou não supervisionada durante a etapa de limpeza dos dados.

Nesse estudo, foram mantidos todos os registros dos IDH de todos os países, mesmos os aparentes *outliers*, pois estes não foram considerados erros, mas valores surpreendentes e corretos, díspares do padrão das demais séries temporais.

Apesar do uso da API não requerer uma pré-análise detalhada das séries temporais, foram realizados outros testes para melhor caracterizar os dados para possíveis comparações com estudos futuros. Entre eles, destacam-se os testes de autocorrelação e correlação cruzada nas séries temporais de um a cinco períodos de defasagem. Nos resultados dos testes, observou-se moderada autocorrelação ($0,33<r<0,66$) em um período de defasagem nas séries temporais univariadas e, a partir de dois períodos de defasagem, a maioria das séries temporais apresentavam autocorrelação baixa ($r<0,33$).

Ao final do pré-processamento de MD, foram selecionadas 376 séries temporais, separadas em dois grupos de dados: o primeiro para previsão do IDH 2013, com dados referentes ao período de 1980 a 2012, e o segundo para previsão do IDH 2014, com dados referentes ao período de 1980 a 2013. Cada grupo de dados apresentava um modelo genérico treinado com séries multivariadas correspondentes aos países filiados à ONU e 187 modelos específicos, treinados com séries univariadas correspondente a cada país, resultando assim em 188 modelos por grupo.

Mineração de dados

A etapa de MD consiste na aplicação de um ou vários algoritmos de aprendizado de máquina. Nesta etapa foi configurada a API e também foram

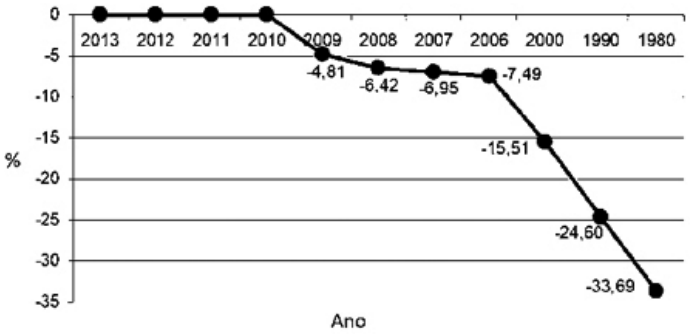

Figura 2. Percentual de ausência de dados temporais.

TABELA III
DESCRIÇÃO ESTATÍSTICA DAS SÉRIES TEMPORAIS CORRESPONDENTES AO IDH

Estatística descritiva	p/Período												p/País
	1980	1990	2000	2005	2006	2007	2008	2009	2010	2011	2012	2013	1980-2013
Média	0,538	0,587	0,620	0,642	0,648	0,656	0,662	0,665	0,670	0,672	0,675	0,686	0,649
Desvio padrão	0,185	0,181	0,188	0,185	0,184	0,182	0,180	0,175	0,172	0,172	0,171	0,156	0,042
Contagem	111	132	152	169	164	166	167	175	187	187	187	187	11,606
NC (95,0%)	0,035	0,031	0,030	0,028	0,028	0,028	0,027	0,026	0,025	0,025	0,025	0,023	0,023

NC: Nível de confiança.

selecionados, testados e configurados os algoritmos disponíveis para mineração.

A API foi configurada considerando-se testes em relação ao número máximo de defasagens e o ajuste ou não da variância. Primeiramente foram realizados testes de previsão nos modelos genéricos (MG) com diferentes configurações de máxima defasagem (2-12). Por meio do teste de análise de variância de Friedman observou-se a inexistência de diferenças significativas ($p=0,61$) em relação ao MAE para diferentes configurações. E com o teste de correlação de Spearman verificou-se alta colinearidade entre o MAE das diferentes defasagens ($0,96<r>1$; $p<0,001$). A partir da interpretação destes resultados, decidiu-se dispensar a definição de valores máximos de defasagem e adotou-se a configuração padrão estabelecida pelo painel de configuração básica da API (lag=5). Em seguida, foram testados os modelos com e sem o ajuste de variância.

Segundo Pentaho (2014) o ajuste de variância pode, ou não, melhorar o desempenho dos algoritmos. No caso específico deste trabalho, verificou-se que o ajuste aumentou o MAE, então decidiu-se não adotá-lo. Ao final destes testes preliminares, foi definida a configuração para a API Forescast do WEKA, conforme apresentado na Tabela IV.

Seleção do algoritmo

Para a seleção do algoritmo mais adequado ao estudo foram testados os algoritmos pertencentes ao grupo de aprendizagem baseada em funções: *least median squared, linear regression, multilayer perceptron*, RBF network, SMOreg (Shevade *et al*., 2000) e *Gaussian processes*. Mantiveram-se suas configurações padrões definidas no ambiente WEKA.

Algumas aplicações de algoritmos foram interrompidas em função do custo operacional elevado, com tempo de resposta >24h para os MGs. Apenas os algoritmos *Gaussian processes* e SMOReg apresentaram custo operacional satisfatório, com tempo de resposta <10min nos MGs. Já nos modelos específicos (ME) todos os algoritmos testados apresentaram tempo de resposta satisfatório, <5s. Também foram realizados testes com retroalimentação de resultados das previsões dos modelos, e verificou-se baixas correlações entre os erros em um horizonte de previsões.

Ao final dos testes preliminares, foi selecionado o algoritmo SMOReg por apresentar os melhores resultados, tanto para os MGs com para os MEs, além de apresentar custo operacional significativamente reduzido em relação aos demais.

O SMOReg aplica um motor de vetores de suporte para regressão. Esta técnica é discutida em Smola e Schölkopf (2004), que também apresentam os algoritmos mais utilizados naquele momento da pesquisa, que inclui o SMO, versão original do SMOReg.

Finalizando a etapa de MD, 376 modelos foram desenvolvidos com o uso do algoritmo SMOreg, dois MGs e 374 MEs, para a previsão do IDH de 187 países nos períodos de 2013 e 2014. Exemplos destes modelos podem ser observados em Santos (2015), que apresenta um ME completo para previsão do IDH do país Chipre e um MG parcial, que na sua forma original é composto de 356.186 linhas.

Pós-processamento de MD

Os resultados dos modelos alimentaram uma base de dados, e através de SQLs específicas, esses resultados foram organizados por modelos e pareados por países. Isto permitiu comparações entre os valores reais disponíveis em UNDATA (2014) e as previsões, assim como entre as medidas de qualidade dos modelos.

A última etapa do processo de KDD, chamada segundo Fayyad *et al*. (1996) de 'conhecimento', que exige a interpretação dos padrões descobertos, será apresentada na seção de resultados, a seguir.

Resultados

As previsões do IDH dos modelos utilizando o algoritmo SMOReg são apresentadas nas Figuras 3 a 6, em que os resultados dos modelos e previsões estão organizados por ordem decrescente do IDH 2013, sendo que cada um agrupa uma categoria de país (desenvolvidos, em desenvolvimento e subdesenvolvidos). Outros detalhes da pesquisa e valores de tendências e previsões do IDH por países podem ser consultados em Santos (2015), que apresenta os dados reais, referentes ao

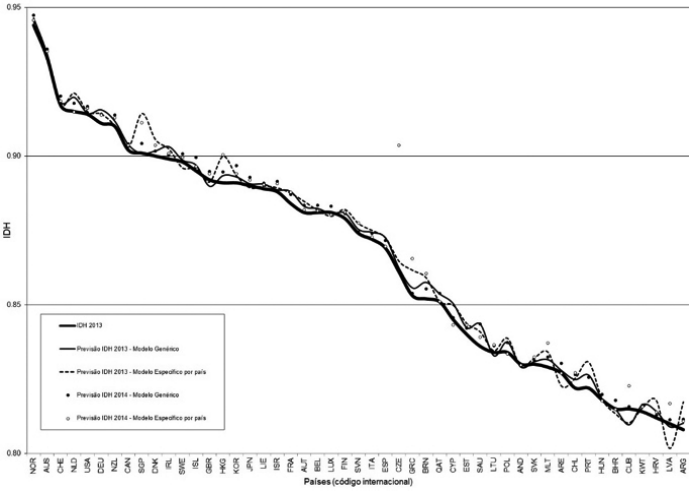

Figura 3. Dados reais (2013) e previsões dos modelos específicos e genéricos para o IDH (2013 e 2014) dos países desenvolvidos (IDH muito alto).

TABELA IV
CONFIGURAÇÃO DA APLICAÇÃO DE INTERFACE DE PROGRAMAÇÃO *Forecast*

Configurações do API *Forecast*				
Básica	Avançadas			
Seleção alvo	Base de aprendizagem (algoritmos)	Criação de defasagem	Avaliação	Saída
Conjunto de dados - Pais(es) Parâmetros Número de vezes de previsão (horizonte de previsão) = 1 Carimbo de tempo <Ano> Periodicidade <Anual> <Ligado> Intervalo de confiança <Ligado> Executar avaliação	*Gaussian processes* *least median squared;* *linear regression;* *multilayer perceptron;* *multilayerperceptron*CS; RBFNetwork; RBFRegressor; SMOReg	<Desligado> Ajustar para variação <Desligado> Uso de comprimentos de defasagens personalizados <Desligado> Seleção de defasagem da sintonização fina	<Ligado> Avaliar durante o treinamento	<Ligado> Saída de previsões futuras no final da série

IDH 2012 e 2013, com as previsões dos IDH 2013 e 2014, a partir dos MGs e MEs, classificados por classes (nível e tipos de IDH) e ranqueados por país em relação ao IDH 2013.

Observa-se na Figura 3, que no grupo de países desenvolvidos, o Reino Unido e Cuba apresentam valores de IDH 2013 acima das expectativas dos modelos. Já no grupo de países em desenvolvimento com IDH elevado (Figura 4) o Peru apresentou valores acima das expectativas, enquanto Líbia, Ilhas Maurício e Belize apresentaram valores inferiores ao estimado. No mesmo grupo, mas com IDH médio (Figura 5), os países África do Sul, Timor-Leste, República Árabe Síria, Zâmbia, São Tomé e Príncipe e Guiné Equatorial, apresentaram valores menores que as previsões.

No grupo de países subdesenvolvidos (Figura 6), Paquistão, Ruanda, Papua Nova Guiné, Afeganistão, Guiné, Burkina Faso, Eritreia, Serra Leoa, República Centro Africano, República Democrá-tica do Congo e Nigéria apresentaram valores inferiores as previsões dos modelos. A Líbia teve a maior queda do IDH (-0,005), cinco posições no ranking, embora as expectativas dos modelos que consideram apenas os dados históricos do IDH era de que seu IDH aumentaria entre 0.8103 a 0.8123. A República Árabe da Síria também

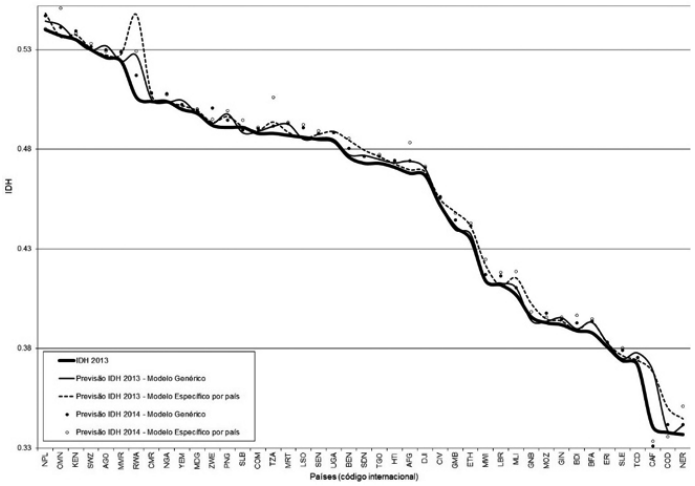

Figura 6. Dados reais (2013) e previsões dos modelos específicos e genéricos para o IDH (2013 e 2014) dos países subdesenvolvidos - IDH baixo.

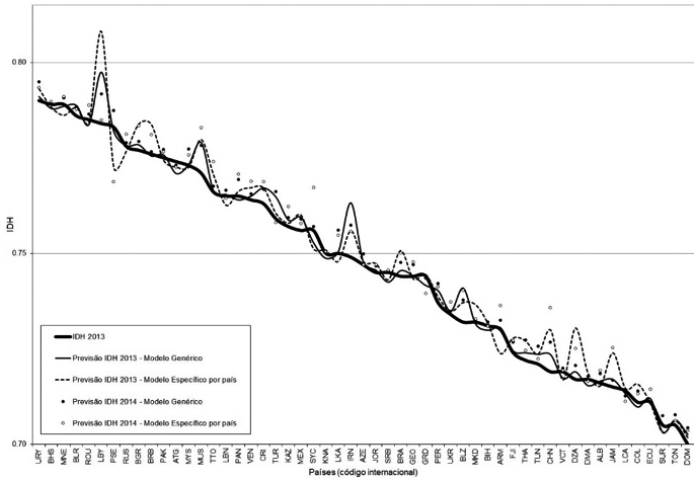

Figura 4. Dados reais (2013) e previsões dos modelos específicos e genéricos para o IDH (2013 e 2014) dos países em desenvolvimento - IDH alto.

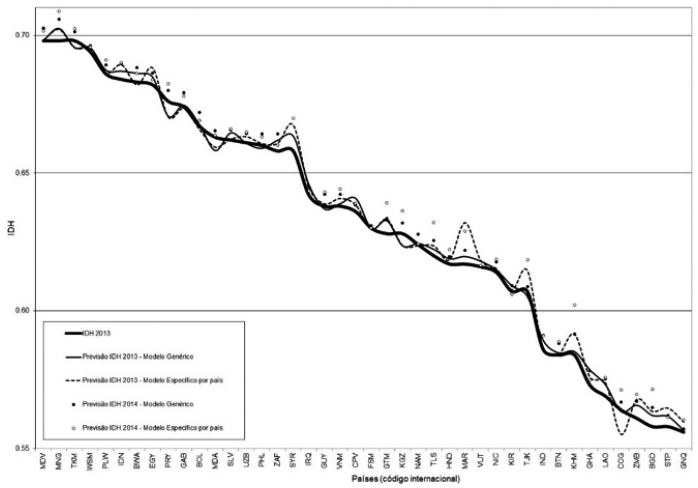

Figura 5. Dados reais (2013) e previsões dos modelos específicos e genéricos para o IDH (2013 e 2014) dos países em desenvolvimento - IDH médio.

apresentou IDH (0,658) menor que o esperado pelos modelos, que seria de 0,6621 a 0,6739. O país apresentou queda no índice (-0.004) bem como queda de quatro posições no ranking. Estas contradições possivelmente ocorreram em consequência da guerra civil que o país vem sofrendo.

O Zimbábue foi o país que mais subiu posições (quatro). No total, dos 187 países avaliados, 35 caíram no ranking e outros 38 subiram. Os demais, 114 países se mantiveram no mesmo nível de desenvolvimento.

Em relação ao valor absoluto do IDH 2013, observou-se que 41,71% dos valores de IDH situaram-se entre os intervalos das previsões dos modelos multivariados e univariados. Ademais, observou-se que 41,18% foram menores simultaneamente que as previsões dos dois modelos, e que 17,11% foram simultaneamente maiores que o previsto pelos modelos. A maioria dos valores absolutos não apresentaram diferenças significativas com as previsões, principalmente em relação ao MG. Verificou-se que 61,50% das previsões com menor erro absoluto correspondiam aos MGs.

Por meio do teste de correlação de Pearson, verificou-se alta correlação (r= 0,99; p<0,0001) das previsões do IDH 2013 com as tendências reais no MG. As previsões deste modelo apresentam capacidade explicativa de 99,6% da variabilidade da tendência do IDH divulgado pela ONU (Figura 7).

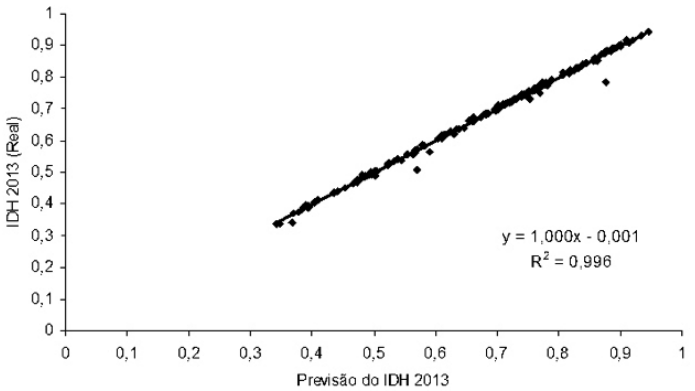

Figura 7. Previsão do IDH 2013 x IDH 2013 real a partir de séries temporais multivariadas.

A Figura 8 apresenta as medidas de qualidade acumuladas das previsões dos MGs e MEs para os anos de 2013 e 2014, com visíveis diferenças de qualidade entre os modelos. Os MGs apresentaram as melhores medidas de qualidade das previsões.

Os testes de análise de variância, T Student e Wilcoxon matched-pairs, apontaram diferenças significativas entre as medidas de qualidade dos modelos ($p<0,001$). Os MGs apresentaram maiores médias da DAC e menores médias de erros que os MEs (Tabela V).

Verificou-se com o teste de correlação de Spearman, a correlação entre as medidas de qualidade das previsões e características estatísticas do IDH, como variabilidade das séries temporais (número de exemplos) e a variância do índice (Tabela VI).

A quantidade de exemplos das séries temporais apresenta baixa correlação com a DAC ($r=0,17$), alta correlação com os tipos de erros ($r>0,66$) nos MGs e moderada correlação ($0,33<r< 0,67$) com os tipos de erros nos MEs.

A variância do IDH apresenta baixa correlação com a DAC ($0,23<r<0,30$), alta correlação com os tipos de erros ($r>0,66$) nos MGs e moderada correlação ($0,33<r<0,67$) como os tipos de erros nos MEs.

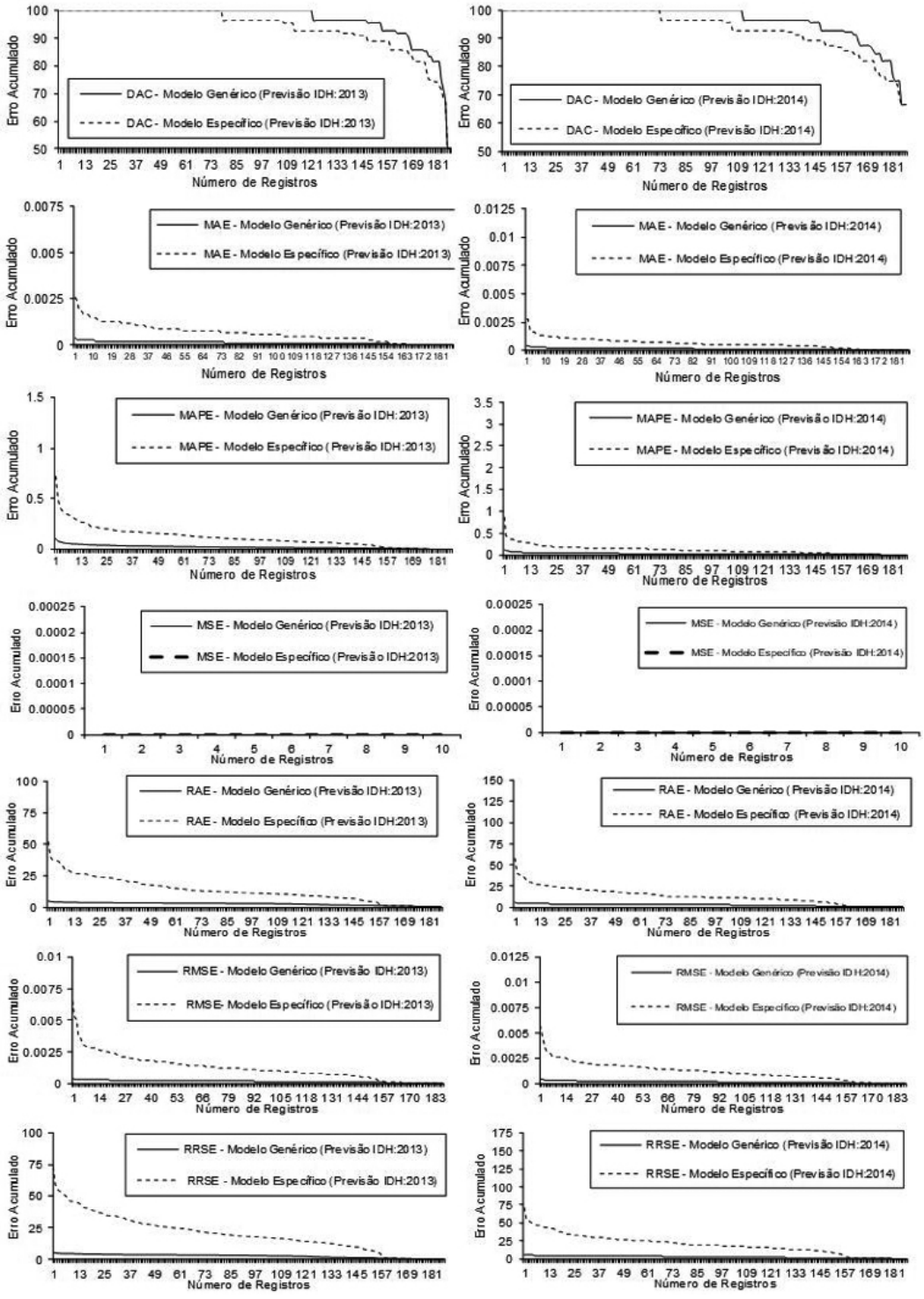

Figura 8. Medidas de qualidades (DAC, MAE, MAPE, MSE, RAE, RMSE; RRSE) das previsões do IDH 2013 e 2014 dos modelos genéricos e dos modelos específicos.

Discussão

O estudo empírico, utilizando dados reais, contribui para a difusão de novos métodos de previsão e complementa o rol de experimentos, que atendem a carência apontada por Keogh e Kasetty (2003), da falta de pesquisas de MD para testar novos métodos.

O experimento fez a previsão e comparação do IDH 2013 com as tendências divulgadas pela ONU em 24 de julho de 2014. Também fez previsões do IDH 2014 que poderão ser confirmadas no relatório seguinte da organização, publicado posteriormente ao término da confecção do presente trabalho.

Segundo Rodrigues e Stevenson (2013), boa parte da literatura sugere que previsões combinadas podem melhorar as previsões individuais. Isto foi visível nos MGs que apresentam melhores resultados que os MEs. Nos MGs o algoritmo aprende com o comportamento de histórico das séries temporais de todos os países, enquanto que nos MEs a aprendizagem se limita às séries temporais do país alvo.

Atualizações significativas dos índices de alguns países podem limitar o estudo. Segundo a UNDP (2014), as estimativas internacionais e nacionais de dados podem apresentar inconsistência, uma vez que as agências de dados internacionais consultam os dados

TABELA V
ESTATÍSTICA DE RESUMO DAS MEDIDAS DE QUALIDADE DAS PREVISÕES DE IDH POR ANO E MODELO

Previsão IDH		2013				2014			
Modelo estatística		Genérico		Específicos		Genérico		Específicos	
		μ	±	μ	±	μ	±	μ	±
Medida de qualidade	DAC	96,52	7,45	91,57	10,10	96,29	6,63	91,15	9,46
	MAE	0,000129	0,000084	0,001475	0,000991	0,000134	0,000084	0,001601	0,001217
	MAPE	0,023444	0,017070	0,258472	0,210777	0,024194	0,018624	0,287996	0,317034
	MSE	0,000000	0,000000	0,000003	0,000016	0,000000	0,000000	0,000004	0,000023
	RAE	2,55	1,18	27,17	15,85	2,67	1,23	29,32	18,30
	RMSE	0,000144	0,000089	0,002436	0,001753	0,000145	0,000089	0,002651	0,001965
	RRSE	2,50	1,19	37,38	20,12	2,62	1,25	40,09	22,09

TABELA VI
CORRELAÇÃO ENTRE AS MEDIDAS DE QUALIDADE DAS PREVISÕES E AS MEDIDAS ESTATÍSTICAS DAS SÉRIES TEMPORAIS: NÚMERO DE ELEMENTOS, VARIÂNCIA DO IDH

Medida de qualidade	Número de exemplos		Variância do IDH	
	Modelos			
	Genérico	Específicos	Genérico	Específicos
DAC	0,17	-	0,23	0,30
MAE	0,67	0,34	0,90	0,40
MAPE	0,67	0,35	0,86	0,39
MSE	-			
RAE	0,8	0,37	0,72	-
RMSE	0,69	0,29	0,92	0,37
RRSE	0,81	0,33	0,76	-

nacionais, e eventualmente estimam dados inexistentes para comparação entre países.

A má qualidade dos dados é um problema que prejudica a MD. Em grandes bases de dados, a ocorrência de erros e dados incompletos é comum (Witten e Frank, 2005). Intervalos de previsões são muitas vezes sensíveis a *outliers*, principalmente quando da ocorrência na proximidade da origem da previsão (Chen e Liu, 1993b). Técnicas de MD em alguns casos podem solucionar alguns desses problemas (Witten e Frank, 2005). A redução de dimensionalidade é uma alternativa que pode ser utilizada para eliminação de ruídos ou dados irrelevantes (Tan *et al*., 2005).

O estudo fez a previsão com 34,22% dos países apresentando dados incompletos retroativos a 2010. Esta ausência representava 12,64% das séries temporais nos MGs e até 69,23% nos MEs (μ= 10,78 ±18,54%). A ausência de dados foi tratada de forma não supervisionada. Os algoritmos utilizados superam essas ausências através de interpolação de dados, principalmente porque as ausências não ocorreram próximas às origens das previsões. Uma pequena interferência da ausência de dados foi perceptível com a diminuição da DAC, principalmente nos MEs.

Em relação ao método utilizado, pesquisas empíricas de Keogh e Kasetty (2003) encontraram pouco ganho com a MD na época. Para Armstrong (2006), os métodos promissores precisam ser replicados para se identificar em que condições estes podem falhar. Em seu estudo, também menciona que as técnicas de MD oferecem pouca promessa, e que talvez a grande falha desses métodos esteja na falta de conhecimento do domínio. Já no presente estudo, foram percorridas todas as etapas do processo de KKD, definidas por Fayyad *et al*. (1996), e isto corrobora com o seu estudo, bem como o estudo de Michalski e Kaufman (1998), que mencionam que a eficácia do método depende do rigor deste processo e que todas as suas etapas são importantes para o produto final, e ainda, que a MD é apenas uma das suas etapas.

Apesar de as técnicas de MD não exigirem o conhecimento prévio do domínio como menciona Armstrong (2006), observou-se no presente estudo que se pode conhecer muito sobre o domínio na etapa de KDD de pré-processamento, que antecede a etapa de MD, principalmente na subetapa de 'exploração da base de dados', como sugere Fayyad *et al*. (1996). E também reafirma-se, a partir da experimentação realizada, que essas etapas são de fundamental importância para condução do processo e para o estabelecimento da técnica mais adequada para o tipo de problema e natureza dos dados (Fayyad *et al*., 1996).

Apesar dos algoritmos de aprendizado de máquina serem recomendados para aquisição de conhecimento, por reduzirem a necessidade de especialistas (Arinze, 1994), a literatura recomenda interação entre especialistas em MD e do domínio investigado (Gargano e Raggad, 1999; Hong e Han, 2002; Kopanas *et al*., 2002; Nemati et al., 2002; Hofmann e Tierney, 2003; Dubey *et al*., 2014; Kadhim *et al*., 2014). Esta interação foi possível no presente estudo, tendo contribuído para melhor compreensão dos dados, bem como dos resultados obtidos.

Apesar de este trabalho testar exclusivamente algoritmos de aprendizagem baseada em funções e comparar o desempenho mais promissor em MG e ME, considerando as características específicas dos dados, observa-se grande avanço na qualidade das previsões obtidas, contrariando estudos anteriores de Keogh e Kasetty (2003) e Armstrong (2006), que não verificaram vantagens no uso das técnicas de MD. O mesmo se observa em outros estudos recentes, como de Lloyd (2014), que apesar de não deixar explícito o uso do processo de KDD, destaca algumas etapas do processo e o uso de técnicas de MD utilizadas para solução do problema de previsão. Ainda, Hong *et al*. (2014) apresentam diversos aspectos da GEFCom2012, incluindo os métodos utilizados pelos participantes, confirmando que algumas técnicas de MD têm vantagens sobre outras populares, como a ARIMA.

Em relação ao custo operacional, segundo Mannila (1996), as etapas mais dispendiosas de tempo são as que antecedem a de MD, podendo consumir até 80% de todo o processo de KDD. Este estudo reafirma os custos operacionais, com aproximadamente 60% do tempo gasto no pré-processamento dos dados, 10% na etapa de MD e 30% no pós-processamento dos dados.

A redução de dimensionalidade pode diminuir custos operacionais e pode eliminar ruídos (Tan *et al*., 2005), aumentando a taxa de acerto. No

entanto, a alta dimensionalidade neste estudo em específico, não aumentou o custo operacional, pois se testou a possibilidade de redução dimensionalidade, mas verificou-se que não houve melhorias significativas do tempo de resposta e da DAC. Já a definição do algoritmo, foi determinante no custo operacional, pois alguns algoritmos ultrapassavam o tempo de processamento esperado, e foram abortados da análise.

Entre as medidas de qualidade das previsões, geralmente o MSE é mais utilizado por resultar em valores na mesma escala dos dados. O RMSE e MSE são muito populares, principalmente porque são muito empregados em modelagens estatísticas (Hyndman e Koehler, 2006), mas são mais sensíveis a *outliers* que outras medidas, como o MAE (Passari, 2003; Hyndman e Koehler, 2006). Armstrong (2001) apresenta uma lista de 32 princípios para avaliar sistematicamente o método de previsão, não recomendando medidas sensíveis a *outliers*. A MAPE é sugerida como a melhor medida neste caso, por ser uma medida absoluta em porcentagem do valor previsto, além de possibilitar uma visão da amplitude do erro (Abraham e Chuang, 1989; Passari, 2003). Já no caso de modelos que respeitem limites de erro máximo, o MAE é o mais indicado. Tanto no MAE como no MSE, durante o somatório, um erro positivo não é anulado por um erro negativo ou vice-versa (Passari, 2003). Peña e Sánchez (2007) utilizam o MSE para apresentar vantagens nos MGs em relação aos modelos univariados, criando uma equação de previsibilidade da série temporal na adoção de preditores multivariados ao invés de univariados.

Estudos como de Greer (2003), para previsões direcionais de taxas de juros de longo prazo, e de Tang *et al*. (2014), utilizam a DAC em seus estudos, a qual fornece a correção da direção prevista, e também pode ser utilizada para avaliar a precisão da previsão. Quanto maior o seu valor, melhores serão as previsões (Wang, *et al*., 2012). MAPE permite comparar modelos com dados diferentes (Passari, 2003). A competição de Energia Global de Previsão de 2012 (GEFCOM, 2012), utilizou o RMSE para avaliar os melhores modelos apresentados na competição. O RMSE permite retornar a medida original dos dados a partir da raiz do MSE (Passari, 2003). Neste estudo, foram utilizadas todas as medidas de avaliação das previsões disponíveis na API, observando-se suas propriedades discutidas na literatura. No entanto, dadas as características do experimento, como utilização de variáveis com uma única unidade de medida, observa-se que apenas as medidas de qualidade DAC e MAE são suficientes para avaliação das previsões do IDH.

Em relação à eficiência do modelo, Putsis (1998) e Lawrence *et al*. (2000) apontam que certas características do erro devem ser observadas. Segundo Putsis (1998), em um modelo eficiente não deve existir correlação entre os erros de um período para outro, o que indica que o modelo aprende com os erros do passado. Esta premissa foi observada nas duas categorias de modelos apresentados neste estudo, pois apesar da alta correlação entre as séries temporais, observou-se que não existiam correlações entre os erros das previsões e entre erros de pontos de observações subsequentes. Para Lawrence *et al*. (2000), a distribuição dos erros deve ter uma forma próxima da normal. Isto também foi verificado no presente estudo, pois os valores dos erros no treinamento dos modelos foram submetidos e aprovados no teste KS. No entanto, esta condição apontada por Lawrence *et al*. (2000) também dependerá do número de exemplos da série temporal.

Os MGs apresentaram melhor desempenho que os MEs. No entanto, esta vantagem relativa do preditor multivariado pode ser muito diferente em cada país. Peña e Sánchez (2007) também destacam vantagens dos MGs, principalmente se existirem fortes relações entre as séries temporais, o que também ocorreu no presente estudo.

A análise de variância apontou diferenças significativas entre os MGs e MEs para a previsão do IDH 2013. O mesmo não ocorreu entre os modelos para previsão do IDH 2014, que não apresentaram diferenças significativas entre si. As previsões confirmadas, referentes ao IDH 2013 apontaram que os resultados dos MGs não apresentam diferenças significativas dos valores reais, e tendem a se aproximar mais destes que os resultados dos MEs.

Na ocasião do término da redação do presente manuscrito, os dados do IDH 2014 ainda não haviam sido divulgados. Portanto, a confirmação ou refutação das tendências apresentadas na presente investigação será realizada em estudo subsequente.

A eficiência dos MGs pode ser explicada implicitamente pelas interdependências e vulnerabilidade dos países apontadas pelo UNDP (2014).

Conclusão

As previsões de eventos com probabilidades de ocorrerem com base em históricos de séries temporais multivariadas ou univariadas são cada vez mais comuns.

Modelos desenvolvidos a partir de séries temporais multivariadas, apesar de mais complexos, se demonstraram mais precisos do que os modelos desenvolvidos a partir de séries univariadas, principalmente se existir alta correlação entre as séries temporais.

As séries temporais multivariadas possibilitam maior aprendizagem dos algoritmos com o aumento de diferentes experiências históricas univariadas.

A execução da MD respeitando todas as etapas do processo de KDD resultou em previsões de séries temporais com precisão satisfatória.

O IDH é um índice robusto com grande previsibilidade e vulnerabilidade. As contradições entre a previsão e os valores reais do índice podem desencadear reflexões e auxiliar em tomadas de decisões para sustentação ou mudanças políticas e econômicas, e também, justificar o cenário vivido por um país ou pelo mundo.

AGRADECIMENTOS

O presente trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil.

REFERENCES

Abraham B, Chuang A (1989) Outlier detection and time series modeling. *Technometrics 31*: 241-248.

Alonso MA, Cruz AV, Barceló G (2009) Pronóstico para la inyección de tenso-activos en pozos de petróleo a partir de uma metodología que integra técnicas de inteligencia artificial y minería de datos. *Interciencia 34*: 703-709.

Arinze B (1994) Selecting appropriate forecasting models using rule induction. *Omega 22*: 647-658.

Armstrong JS (2001) Evaluating forecasting methods. Em Armstrong JS (Eds.) *Principles of forecasting*. Springer. pp. 443-472.

Armstrong JS (2006) Findings from evidence-based forecasting: Methods for reducing forecast error. *Int. J. Forecast. 22*: 583-598.

Chatfield C (1995) Model uncertainty, data mining and statistical inference. *J. Roy. Stat. Soc.* A *158*: 419-466.

Chen C, Liu LM (1993a) Joint estimation of model parameters and outlier effects in time series. *J. Am. Stat. Assoc. 88*(421): 284-297.

Chen C, Liu LM (1993b) Forecasting time series with outliers. *J. Forecast. 12*: 13-35.

Dubey S, Pandey R, Gautam S (2014) Development of multimedia fuzzy based diagnostic expert system for integrated disease management in chickpea. *Int. J. Sci. Mod. Eng. 2*(2): 16-20.

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. *AI Magaz. 17*(3): 37.

Fox AJ (1972) Outliers in time series. *J. Roy. Stat. Soc B 34*: 350-363.

Gargano ML, Raggad BG (1999) Data mining-a powerful information creating tool. *OCLC Syst. Serv. 15*: 81-90.

GEFCOM (2012) Global Energy Forecasting Competition 2012. Wind Forecasting. http://www.kaggle.com/c/GEF2012-wind-forecasting/details/evaluation (Cons. 02/11/2014).

Greer M (2003) Directional accuracy tests of long-term interest rate forecasts. *Int. J. Forecast. 19*: 291-298.

Hofmann M, Tierney B (2003) The involvement of human resources in large scale data mining projects. Em *Proc. 1st Int. Symp. on Information and Communication Technologies*. Dublin, Irlanda. pp. 103-109.

Hong T, Han I (2002) Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks. *Expert Syst. Applic. 23*: 1-8.

Hong T, Pinson P, Fan S (2014) Global energy forecasting competition 2012. *Int. J. Forecast. 30*: 357-363.

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. *Int. J. Forecast. 22*: 679-688.

Kadhim MA, Alam MA, Kaur H (2014) A multi-intelligent agent for knowledge discovery in database (MIAKDD): Cooperative approach with domain expert for rules extraction. Em Huang DS (Ed.) *Intelligent Computing Methodologies*. Vol. 8589. Springer. Suiça. pp. 602-614.

Keogh E, Kasetty S (2003) On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Min. Kknowl. Discov. 7*: 349-371.

Kopanas I, Avouris NM, Daskalaki S (2002) The role of domain knowledge in a large scale data mining project. Em Vlahavas IP, Spyropoulos CD (Eds.) *Proc. 2nd Hellenic Conf. on AI: Methods and Applications of Artificial Intelligence*. Springer. Alemanha. pp. 288-299

Lawrence M, O'Connor M, Edmundson B (2000) A field study of sales forecasting accuracy and processes. *Eur. J. Operat. Res. 122*: 151-160.

Lloyd JR (2014) GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *Int. J. Forecast. 30*: 369-374.

Mangalova E, Agafonov E (2014) Wind power forecasting using the k-nearest neighbors algorithm. *Int. J. Forecast. 30*: 402-406.

Mannila H (1996) Data mining: machine learning, statistics, and databases. *Proc. 8th Int. Conf. on Scientific and Statistical Database Management*. IEEE Computer Society. USA. pp. 2-9.

Michalski RS, Kaufman KA (1998) Data mining and knowledge discovery: A review of issues and a multistrategy approach. Em *Machine Learning and Data Mining: Methods and Applications*. Wiley. pp. 71-112.

Muirhead CR (1986) Distinguishing outlier types in time series. *J. Roy. Stat. Soc.* B *48*: 39-47.

Nemati HR, Steiger DM, Iyer LS, Herschel RT (2002) Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decis. Supp. Syst. 33*: 143-161.

Palit AK, Popovic D (2006) *Computational intelligence in time series forecasting: theory and engineering applications*. Springer. 371 pp.

Passari AFL (2003) *Exploraçao de Dados Atomizados para Previsao de Vendas no Varejo Utilizando Redes Neurais*. Tese. Universidade de São Paulo. Brasil. 143 pp.

Peña D, Sánchez I (2007) Measuring the advantages of multivariate vs. univariate forecasts. *J. Time Ser. Anal. 28*: 886-909.

Pentaho (2014) *Time Series Analysis and Forecasting with Weka*. http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka (Cons. 02/11/2014).

Putsis WP (1998) Parameter variation and new product diffusion. *J. Forecast. 17*: 231-257.

Rodrigues BD, Stevenson MJ (2013) Takeover prediction using forecast combinations. *Int. J. Forecast. 29*: 628-641.

Santos CB (2015) *Pesquisa. Previsão de Séries Temporais. Previsão do IDH 2013 e 2014*. https://sites.google.com/site/celsobilynkievyczdossantos/home/pesquisas/TabelaResultadosPrevis%C3%A3oIDH2013e2014.htm?attredirects=0&d=1 (Cons. 15/04/2015).

Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK (2000) Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Netw. 11*: 1188-1193.

Silva L (2014) A feature engineering approach to wind power forecasting: GEFCom 2012. *Int. J. Forecast. 30*: 395-401.

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. *Stat. Comput. 14*: 199-222.

Tan PN, Steinbach M, Kumar V (2005) *Introduction to Data Mining*. Addison-Wesley Longman. Boston, MA, EUA. 978 pp.

Tang L, Yu L, He K (2014) A novel data-characteristic-driven modeling methodology for nuclear energy consumption forecasting. *Appl. Energy 128*: 1-14.

UNDATA (2014) *Human Development Index trends*, 1980–2013. United Nations Development Programme. http://data.un.org/DocumentData.aspx?id=364 (Cons. 02/11/2014).

UNDP (1990) *Human Development Report (HDR) 1990: Concept and Measurement of Human Development*. United Nations Development Programme. Press OU. Nova York, EUA.

UNDP (2014) *Human Development Report (HDR) 2014. Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience*. United Nations Development Programme. Press OU. Nova York, EUA.

Wang JJ, Wang JZ, Zhang ZG, Guo SP (2012) Stock index forecasting based on a hybrid model. *Omega 40*: 758-766.

Witten IH, Frank E (2005) *Data Mining: Practical Machine Learning Tools and Techniques.* (2ª ed.) Kaufmann. San Francisco, CA, EUA. 558 pp.