Uso de ciência de dados para predição do consumo de fertilizantes no Brasil

Adalberto Andrade[1], Rebecca Salles[1], Flavio Carvalho[1], Eduardo Bezerra[1], Jorge Soares[1], Cristina Gomes de Souza[1], Pedro Henrique Gonzalez[1], Eduardo Ogasawara[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ

{adalberto.andrade,rebecca.salles,flavio.carvalho}@eic.cefet-rj.br

{ebezerra,jorge.soares,cristina.souza}@cefet-rj.br

pegonzalez@eic.cefet-rj.br, eogasawara@ieee.org

Abstract. *Fertilizers are critical elements for food production. Predicting efficiently how fertilizer consumption will behave in the next years allows to properly plan the increase in production and, thereby, mitigate environmental problems resulting from such an increase in production. Given the described elements, this research explores data science approaches to enable the prediction of fertilizer consumption, through the optimization of model construction. The results indicate that the use of the analytical tools presented here may be a way to obtain reliable forecasts to plan future demands.*

Resumo. *Fertilizantes são elementos críticos para à produção de alimentos. Prever eficientemente como o consumo de fertilizantes irá se comportar nos próximos anos permite planejar adequadamente o aumento da produção e, com isso, mitigar problemas ambientais decorrentes de tal aumento de produção. Tendo em vista os elementos citados, esta pesquisa explora abordagens de ciência de dados para capacitar a predição do consumo de fertilizantes, através da otimização da construção de modelos. Os resultados indicam que o uso das ferramentas analíticas aqui apresentadas pode vir a ser uma maneira de obter previsões confiáveis para planejar demandas futuras.*

1. Introdução

Fertilizantes são insumos essenciais na agricultura e sua utilização é a chave para alcançar a segurança alimentar no mundo [Stewart and Roberts, 2012]. Quanto maior for a necessidade de produção de alimentos, maior será a quantidade de fertilizantes necessária para o desenvolvimento da atividade agrícola.

De acordo com a Organização das Nações Unidas [UN, 2019], a população mundial atingirá 9,8 bilhões de pessoas até 2050. Assim, a produção de alimentos terá de aumentar quase 50% acima dos níveis atuais para acompanhar o crescimento da demanda [FAO, 2019]. Para que esse aumento seja possível, sem que haja comprometimento ambiental [Attallah et al., 2019], o planejamento adequado da expansão da produção de fertilizantes é essencial [FAO, 2019].

Esta pesquisa explora diferentes abordagens analíticas, através da otimização da construção de modelos, para realizar a predição do consumo de fertilizantes e com isso permitir um planejamento adequado que possa levar a uma redução no impacto causado ao meio ambiente pelo aumento da produção dos principais fertilizantes. As abordagens são

avaliadas utilizando quatro diferentes séries temporais, referentes ao consumo do quatro fertilizantes mais utilizados no Brasil (N, P_2O_5, K_2O, e NPK).

Além desta introdução, este artigo está organizado em mais quatro seções. A Seção 2 apresenta os principais trabalhos relacionados. A Seção 3 detalha a metodologia desenvolvida nesta pesquisa. Já a Seção 4 apresenta a avaliação experimental e a discussão conduzida neste trabalho. Finalmente, a Seção 5 conclui e apresenta possibilidades de trabalhos futuros.

2. Trabalhos relacionados

A relação de artigos foi obtida por meio de um mapeamento sistemático da base indexada SCOPUS, utilizando a *string* ("predict" OR "forecast") AND "fertilizer" AND ("consumption" OR "demand"), no dia 2 de outubro de 2019. A Tabela 1 destaca trabalhos em periódicos e conferências em língua inglesa sobre predição do consumo de fertilizantes em países ou no mundo, obtidos através da busca supracitada.

Tabela 1. Trabalhos relacionados sobre predição do consumo de fertilizantes

Artigo	Região	Fertilizante	Domínio	Metódo
Styhr Petersen [1977]	Dinamarca	N	Agricultura	Regressão
Deadman and Ghatak [1979]	Mundo	N, P, K	Agricultura	Regressão
Gilland [1993]	Mundo	N	Agricultura	Regressão
Howarth et al. [2002]	EUA	N	Ambiental	Regressão
Dobermann and Cassman [2005]	Mundo	N	Agricultura	Regressão
Zhang and Zhang [2007]	Mundo	N, P, K	Ambiental	Regressão
Tenkorang and Lowenberg-Deboer [2009]	Mundo	N, P, K	Agricultura	Regressão
Ogasawara et al. [2013]	Brasil	NPK, S, N, P, K	Agricultura	ARIMA
Pires et al. [2015]	Brasil	N	Agricultura	Regressão

Modelos causais, tais como modelos de regressão linear/polinomial, são por vezes preferíveis quando estão disponíveis projeções de variáveis relacionadas [Verma and Pearl, 1991]. Assim sendo, a maioria destes artigos estudados faz uso de modelos causais para prever o consumo de fertilizantes, que associam uma ou mais variáveis de interesse à geração de uma curva característica da sua relação.

3. Metodologia

Buscando prever o aumento da demanda para que seja possível mitigar o impacto ambiental causado pelo aumento da produção de fertilizantes, propõe-se neste trabalho a metodologia `opt`. Esta metodologia se resume em quatro etapas, conforme definida na Figura 1 e descrita imediatamente abaixo:

A. Seleção dos Dados e Limpeza: São acessados dados públicos da *International Fertilizer Association* (IFA) [1] e selecionados dados dos quatro fertilizantes mais consumidos no Brasil, de 1961 a 2016. São organizadas séries temporais anuais de nitrogênio (N), fosfato (P_2O_5), potássio (K_2O) e nitrogênio-fosfato-potássio (NPK). Cada série temporal possui 56 observações. A limpeza corresponde ao

[1] Disponível em http://www.fertilizer.org

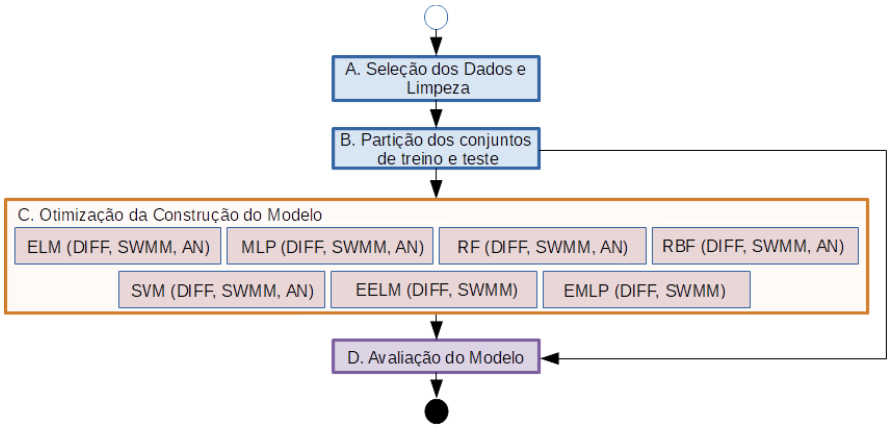


Figura 1. *Workflow* analítico de dados aplicado na metodologia

tratamento de dados faltantes. Nesta pesquisa, dados faltantes no início de uma série temporal foram descartados, como é o caso da Rússia que só tem dados a partir de 1990. Enquanto que dados faltantes no meio de uma série temporal foram interpolados, tal como ocorre com o Paquistão.

B. Partição dos conjuntos de treino e teste: A série temporal anual é tratada como uma série temporal t_i com n observações. Sendo o objetivo prever k observações passo à frente e considerando a ordenação dos dados, a série temporal é dividida de t_1 a t_{n-k} para treino-validação e t_{n-k+1} a t_n para testes.

C. Otimização da Construção do Modelo: Treino-validação é ainda particionado em treino e validação. Essa divisão é replicada r vezes. O treino do modelo real usa observações de t_1 a t_{n-2k-p} e a validação usa de $t_{n-2k+1-p}$ a t_{n-k-p}, para cada replicação definida por p, de modo que $p \in \{0, \cdots, r-1\}$. O processo de replicação otimiza os hiperparâmetros e identifica a melhor abordagem utilizando *grid search* para obtenção de modelos estáveis [Thornton et al., 2013], i.e. um par de pré-processamento de dados e de aprendizado de máquina. Das possíveis opções de modelos, cada série temporal é treinada por 168 modelos diferentes, multiplicado por (r) replicações na otimização da construção do modelo. A Tabela 2 descreve as dimensões do espaço de hiperparâmetros que consideramos como opções. As opções de pré-processamento de dados avaliadas são diferenciação (diff) [Salles et al., 2019], janela deslizante com normalização min-max (swmm) [Han et al., 2011] e normalização adaptativa (an) [Ogasawara et al., 2010]. Os métodos de aprendizado de máquina utilizados são *Extreme Learning Machine* (elm) [Zhao et al., 2016], *Multilayer Perceptron* (mlp) [Coulibaly et al., 2001], *Random Regression Forest* (rf) [Palmer et al., 2007], *Random Base Function* (rbs) [Sfetsos and Coonick, 2000], *Support Vector Machine* (svm) [Sapankevych and Sankar, 2009], *Ensemble Extreme Learning Machine* (eelm) [Zhang and Berardi, 2001], *Ensemble Multilayer Perceptron* (emlp) [Zhang and Berardi, 2001]. A estrutura interna varia de acordo com o tipo de aprendizado de máquina: (i) elm e mlp (o número de nós ocultos); (ii) rf (o número de árvores de decisão); (iii) svm (o tipo de núcleos utilizados: linear, polinomial, base radial e sigmóide); (iv) eelm (o número de eelm interno para o conjunto) e (v) emlp (o número de emlp interno para o conjunto). Por último define-se o número de entradas, que varia de 3 a 10.

D. Avaliação do Modelo: Uma vez obtido o modelo da etapa de otimização da

Tabela 2. (C) Otimização da construção do modelo

Opções de modelos	Valores dos candidatos
Pré-processamento de dados	{diff, swmm, an}
Aprendizado de máquina	{elm, mlp, rf, rbs, svm, eelm, emlp}
Número de entradas	[3..10]
Estrutura interna	{nós ocultos, árvore de decisão, núcleos (linear, polinomial, base radial, e sigmóide), eelm interno, emlp interno}

construção do modelo, ele é utilizado para prever observações para o conjunto de testes. Para fazer isso, é feito treinamento adicional usando todo o conjunto de validação de treinamento para prever k passos à frente de observações. A predição é comparada ao conjunto de testes. Para a análise do erro, utiliza-se o operador SMAPE [Hyndman and Koehler, 2006]. Além disso, a predição é comparada contra a predição usando o modelo `arima` ajustado com o algoritmo Hyndman-Khandakar [Hyndman and Khandakar, 2008]. Esse processo avalia a qualidade do modelo de predição [Salles et al., 2015].

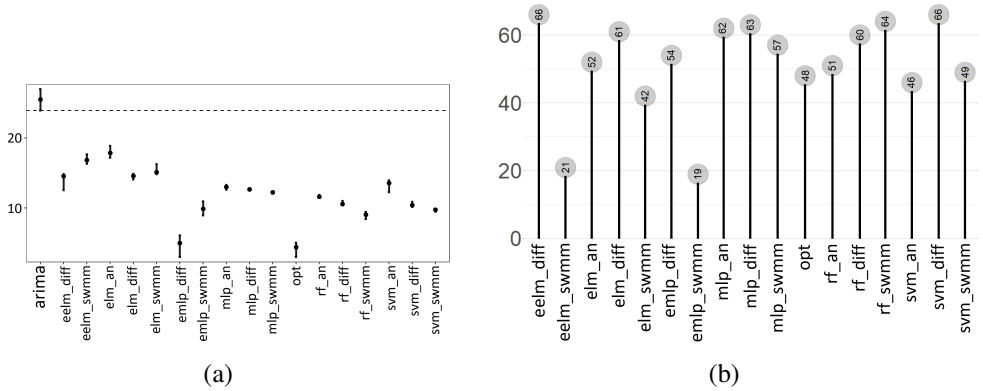
4. Experimentos

A avaliação experimental realiza predições de k passos à frente para o consumo de fertilizantes do Brasil, utilizando quatro séries temporais anuais dos fertilizantes N, P_2O_5, K_2O e NPK. São utilizados dados de 1961 a 2008 para treino e de 2009 a 2016 para testes. Desta forma, por exemplo, ao definir o valor de $k = 1$, o objetivo é prever o consumo de fertilizantes de 2009, enquanto que para $k = 8$, o objetivo é prever, de uma só vez, de 2009 a 2016, variando o valor de replicação (r) de 1 a 5. Na comparação das predições, `opt`, `arima` e todas as abordagens são testadas. Nas Figuras 2.a, 2.b, 3.a, 3.b e 4.a o desempenho de `opt`, `arima` e todas as abordagens são calculados a partir da soma SMAPE gerado na predição de cada série temporal. Na Figura 4.b, o desempenho é apresentado com o valor SMAPE por série temporal das duas abordagens selecionadas (`emlp_diff` e `eelm_diff`) e `opt`, sendo ainda tais valores comparados com o valor do `arima`.

A Figura 2.a apresenta o desempenho da predição de cada abordagem nas quatro séries temporais na fase de treino-validação. Para cada método testado, a barra corresponde ao valor da mediana do SMAPE [Hyndman and Koehler, 2006] com seu intervalo de confiança. Quanto menor o valor do SMAPE, melhor é a predição, e, de acordo com este resultado, `opt` apresentou um desempenho superior a todas abordagens, enquanto `emlp_diff` obteve o segundo melhor desempenho.

Já em relação a fase de teste, a Figura 2.b apresenta a porcentagem de vezes que cada abordagem ganhou do `arima`. Seguindo os resultados do treino, `emlp_diff` foi superior ao `arima`, no entanto, outras abordagens como `eelm_diff` e `svm_diff` que se saíram não tão bem no treino, se mostraram superiores na fase de teste. Além disso, a técnica que obteve melhor resultado na fase de treino, `opt`, se mostrou pior do que o `arima` na fase de teste. Esse resultado exemplifica a dificuldade da tarefa que este trabalho se propõe a lidar.

A Figura 3.a apresenta a diferença percentual na fase treino-validação entre as principais abordagens e `arima` para predições passo à frente. Considerando que valores percentuais acima de zero indicam desempenhos de predições melhores do que `arima`.

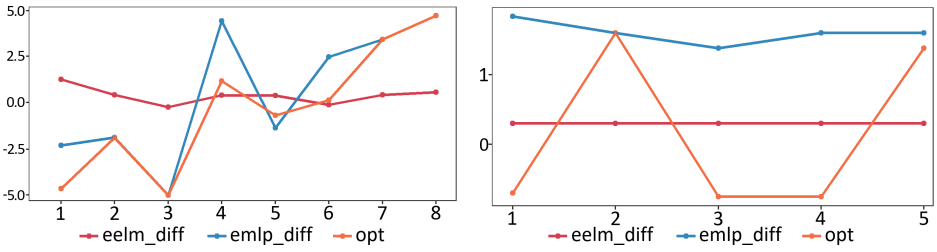


(a) (b)

Figura 2. (a) Medida da percentagem de erro geral (SMAPE) para predição do consumo de fertilizantes das diversas abordagens na etapa de treino-validação (b) Medidas percentuais durante os testes em que cada abordagem teve um desempenho superior ao `arima`

No geral, `opt` e `emlp_diff` aumentam o seu desempenho em relação ao `arima` à medida que o passo à frente aumenta. Enquanto que o desempenho de `eelm_diff` não aumenta à medida que o passo à frente aumenta.

A Figura 3.b apresenta como a replicação em treino-validação altera a predição do modelo durante os testes. O não uso da replicação (valor igual a 1) levou `emlp_diff` a um melhor desempenho. Com a adoção da replicação (valor maior que 1) não fez diferença para `eelm_diff`, enquanto que para `opt`, após oscilação, houve uma melhora em seu desempenho.



(a) (b)

Figura 3. Diferença percentual entre principais abordagens e `arima` para predições passo à frente (a) e replicação (b)

A Figura 4.a apresenta diferença percentual do SMAPE de cada abordagem em relação ao SMAPE do `arima` para predição do consumo de fertilizantes na fase de teste. As barras com valores maiores ou menores que zero, respectivamente, correspondem predições melhores ou piores do que `arima`. Observa-se que `rf_swmm` ficou acima de 3% melhor do que `arima`, embora não tivesse sido escolhida durante treino-validação. Considerando predições por tipos de fertilizantes, a Figura 4.b, apresenta os resultados da Figura 4.a separado por fertilizante para `opt` e os métodos `emlp_diff` e `eelm_diff` que são, respectivamente, a proposta desse trabalho, o que melhor se saiu no treino (tirando o `opt`) e o que melhor se saiu no teste.

É possível observar que, no caso do K_2O, os métodos obtiveram predições bem

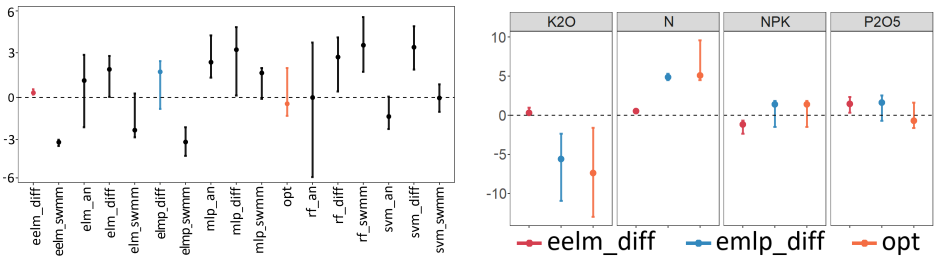


Figura 4. Diferença percentual do SMAPE (`arima` e abordagens) nas predições (a) e diferença para cada fertilizante, usando as abordagens selecionadas (b)

abaixo do `arima`, sendo o `emlp_diff` o único que superou o `arima` sempre. Para o N todas as abordagens conduzem a predições melhores do que `arima`. Já para o NPK, tanto o `eelm_diff`, quanto o `opt`, foram superiores ao `arima` na média. Por último, para o P_2O_5, apenas o `opt` ficou com a média inferior a eficácia do `arima`.

Por último, a Figura 5 compara dados reais (de 2009 a 2016) com predições (8 passos à frente) das principais abordagens e `arima` para cada fertilizante aqui estudado, com replicação de 1 até 5. Se observa que, quanto mais distante o ano a ser previsto, mais distantes da realidade ficaram as predições. No entanto, pode-se constatar também que de modo geral, todos os três métodos analisados se mostraram, na média, superiores ao `arima`.

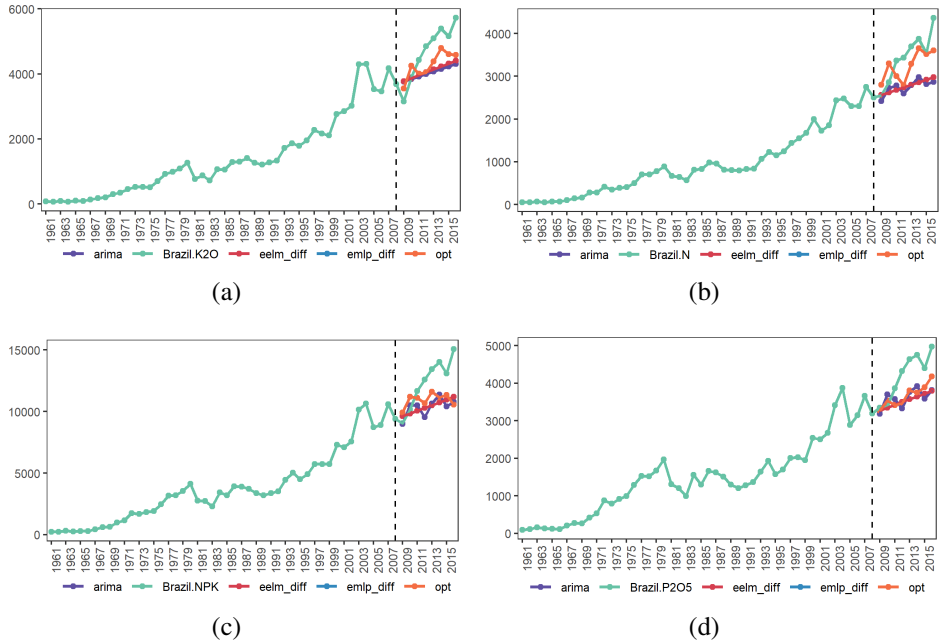


Figura 5. Comparação dos métodos, com predições para oito passos à frente: (a) K_2O, (b) N, (c) NPK e (d) P_2O_5

5. Conclusões

Neste artigo, diversas abordagens de predição são avaliadas em quatro diferentes séries temporais, correspondentes a quatro principais fertilizantes consumidos no Brasil, a fim

de melhorar as predições do consumo de fertilizantes sob diferentes horizontes. Neste sentido, essa pesquisa implementa `opt`, bem como todos as possíveis combinações de métodos do `opt` e os comparamos com o `arima`, usado como modelo base. Na fase de treino `opt` teve desempenho superior a todos os métodos e se mostrou muito superior ao `arima`. Na fase de teste `opt` perde posição frente ao `arima` e aos demais métodos, mas com uma diferença de desempenho menor. Ao final dos experimentos, as abordagens `opt`, `emlp_diff` e `eelm_diff` foram selecionadas por serem as melhores ou no treino, ou no teste, e tiverem sua eficácia analisada para a predição do consumo de cada um dos fertilizantes.

Os resultados indicam que o uso das ferramentas analíticas apresentadas pode vir a ser uma maneira de obter previsões confiáveis para planejar demandas futuras. Mostram também que há espaço para desenvolver formas mais avançadas de seleção de modelos, considerando os resultados do `opt` nos testes. Trabalhos futuros poderão focar no desenvolvimento de métodos de optimização, tais como algoritmos genéticos e outras meta-heurísticas, para melhorar a etapa de seleção de modelo.

Agradecimentos

Os autores agradecem à CAPES (Código de Financiamento 001), FAPERJ e CNPq pelo financiamento parcial desta pesquisa.

Referências

Attallah, M., Metwally, S., Moussa, S., and Soliman, M. A. (2019). Environmental impact assessment of phosphate fertilizers and phosphogypsum waste: elemental and radiological effects. *Microchemical Journal*, 146:789–797.

Coulibaly, P., Anctil, F., and Bobée, B. (2001). Multivariate reservoir inflow forecasting using temporal neural networks. *Journal of Hydrologic Engineering*, 6(5):367–376.

Deadman, D. and Ghatak, S. (1979). Forecasting fertilizer consumption and production: Long- and short-run models. *World Development*, 7(11-12):1063–1072.

Dobermann, A. and Cassman, K. (2005). Cereal area and nitrogen use efficiency are drivers of future nitrogen fertilizer consumption. *Science in China. Series C, Life sciences / Chinese Academy of Sciences*, 48 Spec No:745–758.

FAO (2019). Food and agriculture organization of the united nations. Technical report, http://www.fao.org.

Gilland, B. (1993). Cereals, nitrogen and population: an assessment of the global trends. *Endeavour*, 17(2):84–88.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Haryana, India; Burlington, MA, 3 edition.

Howarth, R., Boyer, E., Pabich, W., and Galloway, J. (2002). Nitrogen use in the United States from 1961-2000 and potential future trends. *Ambio*, 31(2):88–96.

Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Ogasawara, E., De Oliveira, D., Paschoal Jr., F., Castaneda, R., Amorim, M., Mauro, R., Soares, J., Quadros, J., and Bezerra, E. (2013). A forecasting method for fertilizers consumption in Brazil. *International Journal of Agricultural and Environmental Information Systems*, 4(2):23–36.

Ogasawara, E., Martinez, L., De Oliveira, D., Zimbrão, G., Pappa, G., and Mattoso, M. (2010). Adaptive Normalization: A novel data normalization approach for non-stationary time series. In *Proceedings of the International Joint Conference on Neural Networks*.

Palmer, D., O'Boyle, N., Glen, R., and Mitchell, J. (2007). Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, 47(1):150–158.

Pires, M., Da Cunha, D., De Matos Carlos, S., and Costa, M. (2015). Nitrogen-use efficiency, nitrous oxide emissions, and cereal production in Brazil: Current trends and forecasts. *PLoS ONE*, 10(8).

Salles, R., Belloze, K., Porto, F., Gonzalez, P., and Ogasawara, E. (2019). Nonstationary time series transformation methods: An experimental review. *Knowledge-Based Systems*, 164:274–291.

Salles, R., Bezerra, E., Soares, J., and Ogasawara, E. (2015). Evaluating Linear Models as a Baseline for Time Series Imputation. In *XXX Simpósio Brasileiro de Banco de Dados*, Petrópolis, RJ.

Sapankevych, N. and Sankar, R. (2009). Time series prediction using support vector machines: A survey. *IEEE Computational Intelligence Magazine*, 4(2):24–38.

Sfetsos, A. and Coonick, A. (2000). Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy*, 68(2):169–178.

Stewart, W. and Roberts, T. (2012). Food security and the role of fertilizer in supporting it. In *Procedia Engineering*, volume 46, pages 76–82.

Styhr Petersen, H. (1977). Forecasting Danish nitrogen fertilizer consumption. *Industrial Marketing Management*, 6(3):211–222.

Tenkorang, F. and Lowenberg-Deboer, J. (2009). Forecasting long-term global fertilizer demand. *Nutrient Cycling in Agroecosystems*, 83(3):233–247.

Thornton, C., Hutter, F., Hoos, H., and Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume Part F128815, pages 847–855.

UN (2019). United nations. Technical report, https://www.un.org/en/.

Verma, T. and Pearl, J. (1991). *Equivalence and synthesis of causal models*. UCLA, Computer Science Department.

Zhang, G. and Berardi, V. (2001). Time series forecasting with neural network ensembles: An application for exchange rate prediction. *Journal of the Operational Research Society*, 52(6):652–664.

Zhang, W. and Zhang, X. (2007). A forecast analysis on fertilizers consumption worldwide. *Environmental Monitoring and Assessment*, 133(1-3):427–434.

Zhao, Y., Ye, L., Li, Z., Song, X., Lang, Y., and Su, J. (2016). A novel bidirectional mechanism based on time series model for wind power forecasting. *Applied Energy*, 177:793–803.