Artificial Neural Networks applied to assess the impact of $PM_{2.5}$ on hospital admissions for cardiovascular diseases

Jéssica C. Santos-Silva[1], Yara S. Tadano[2], Hugo V. Siqueira[3], Sandra H. W. Medeiros[4], Luiz V. Silva[4], Danielli V. Ferreira[4], Thomas S. Pereira[2], Carlos I. Yamamoto[5], Ricardo H. M. Godoi[1]

[1]*Postgraduate Program in Water Resources and Environmental Engineering, Federal University of Parana*
Curitiba, 81530-000, Paraná, Brazil
jessica.jcss@gmail.com, rhmgodoi@ufpr.br
[2]*Postgraduate Program in Mechanical Engineering, Federal University of Technology - Parana*
Ponta Grossa, 84017-220, Paraná, Brazil
yaratadano@utfpr.edu.br
[3]*Department of Electric Engineering, Federal University of Technology - Parana*
Ponta Grossa, 84017-220, Paraná, Brazil
hugosiqueira@utfpr.edu.br
[4]*Department of Environmental and Sanitary Engineering, University of the Region of Joinville*
Joinville, 89219-710, Santa Catarina, Brazil
sandra.westrupp@gmail.com, luizvitordasilva@gmail.com, danielli.venttura@gmail.com
[5]*Department of Chemical Engineering, Federal University of Paraná, Curitiba, Paraná, Brazil*
Curitiba, 81530-000, Paraná, Brazil
citsuo@gmail.com

Abstract. The high emission of atmospheric pollutants in large urban centers causes several harms to population health. Thus, it is necessary to evaluate the negative impact of its concentration to assist in decision-making and public policies by government agents. Several modeling techniques have been used to assess the effects of air pollution on human health. However, due to their greater flexibility in analyzing the complex nonlinearity of environmental data, Artificial Neural Networks (ANN) have been shown to be the most attractive approach for solving such data modeling problems. This work aimed to compare the performance of two artificial neural networks, Multilayer Perceptron (MLP) and Extreme Learning Machine (ELM) in estimating the number of hospital admissions for cardiovascular diseases due to the concentration of fine particulate matter ($PM_{2.5}$) in Joinville, Brazil. Daily $PM_{2.5}$ concentration and meteorological variables were considered as input variables. MLP network was able to achieve better performance to estimate hospital attendance because of these environmental conditions after three days of $PM_{2.5}$ exposure. The results demonstrate that ANN can be used to predict hospital admissions due to air pollution levels or adverse meteorological conditions and therefore, be used to guide government public policies on air quality and health risk assessment.

Keywords: extreme learning machines; multilayer perceptron; particulate matter; prediction.

1 Introduction

Cities are well-known centers of pollution sources as they shelter 55% of the world's population and 85% of its economic activities [1]. The high impact of air pollutants on human health is considered a dangerous reality that affects 99% of the world population [2]. Ambient exposure to $PM_{2.5}$ has been estimated to cause more than 4 million premature deaths worldwide in 2016, and 118 million lost Disability-Adjusted Life Years (DALYs), being the main driver of air pollution's burden disease worldwide [3], [4]. Therefore, as a public health problem, understanding the association between air pollution and adverse health effects is essential to control and manage its risks.

To better understand the influence of air pollution on the population's health several modeling techniques such as Generalized Linear Models (GLM) and Generalized Additive Models (GAM) have been used to assess the effects of air pollution on human health [5]–[11]. However, when analyzing the complex nonlinearity of environmental data, the greater flexibility of Artificial Neural Networks (ANN) has shown them as the most attractive approach for solving such data modeling problems, even more in developing countries where the lack of data is a reality, as is the case in most Brazilian cities. The ANN have an intrinsic learning capability and generalization ability [12]–[14], and consequently, are important candidates to solve mapping problems [15].

In this work, we compared the performance of two artificial neural networks, Multilayer Perceptron (MLP) and Extreme Learning Machines (ELM) in estimating the number of hospital admissions due to air pollution exposure. The MLP is the most known and frequently used artificial neural network, while ELM are relatively new and, as unorganized machines, their adjustment is simple and fast, requiring only a short time to be trained [11], [15]. The goal of this study is to forecast morbidity from cardiovascular diseases, considering variables regarding air pollution in an urban-industrial region in the municipality of Joinville, in southern Brazil.

2 Material and methods

In this section, the considered problem; data collection and characterization; artificial neural networks characteristics; and computational details were presented.

2.1 Considered problem

The case study considers the concentration of particulate matter with an aerodynamic diameter less than or equal to 2.5 micrometers ($PM_{2.5}$) and meteorological influence on hospital admissions for cardiovascular diseases in Joinville, an important industrial city in southern Brazil. Joinville is the most populous city in Santa Catarina State, it has about 600,000 inhabitants covering an area of 1,125 km² [16]. To our knowledge, this is the first study related to predictive modeling of air pollution health outcomes in Santa Catarina State. Beyond this, while there have been other studies employing ANN to assess the effects of air pollution on human morbidity and mortality, this is the first to model hospital admissions for cardiovascular diseases associated with exposure to ambient $PM_{2.5}$.

2.2 Data collection and characterization

The input variables are characterized as air pollutants ($PM_{2.5}$ concentration) and meteorological conditions (temperature, relative humidity, and precipitation). The considered output was morbidity for cardiovascular diseases. $PM_{2.5}$ concentrations were collected daily from August 2018 to February 2020. Daily hospital admission data due to cardiovascular health problems (Codes I00-I99 from the International Classification of Diseases – ICD 10) were obtained from the Informatics Department of the Unified Health System (DATASUS) using the *microdatasus* package in R [17]. It is necessary to highlight that the information comprises only public health, disregarding data from health insurance and private morbidity. Meteorological data were obtained from Santa Catarina Civil Defense's meteorological stations' network and from the Airport station (SBJV) available on the MESONET website [18]. The precipitation was obtained from the rain gauge network of CEMADEM [19]. As such database is given in 5-min intervals, we calculated the daily averages. For missing data, homogenization and interpolation were done using the Climatol package in R [20], [21]. Besides that, the temporal variables 'day of the week' and 'holidays' were also included as input, since, like meteorological variables, they are important local confusion factors that control acute changes in air pollutants levels as well as hospital admission patterns.

2.3 Artificial Neural Networks

The basic information-processing unit fundamental to the operation of a neural network is the *artificial neuron*, the most basic structure of an ANN, i.e., the functional units responsible for processing the information and providing the output response [22]. In a layered neural network, the neurons are organized in the form of layers, comprising an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight during the training process. The learning algorithm used to train the network is given by its structure, i.e., the way the neurons are linked. The two ANN architectures used in this study are described below (further details can be found in Haykin [12] and Araújo et al. [15])

Remark 1: Multilayer Perceptron (MLP). MLP is a feed-forward neural network, in which the information flows only in one direction: from the input to the output layer [12], [13]. It is a universal approximator suitable for dealing with static problems since it can approximate any nonlinear function if they are limited, continuous, differentiable, and defined in a compact space [11], [12], [15]. In an MLP, the artificial neurons are distributed in three kinds of layers. From an input layer, the data is transmitted to an intermediate (hidden) layer, where the neuron performs a nonlinear transformation, mapping the input signal to another space. Then, the signal is sent to the output layer, where an output signal is generated based, generally, on a linear combination. As a feedforward model, neurons from the same layer are disconnected while those from disjoint layers exchange information [23]. During training, the adjustment of the weights of the neurons is performed in two phases: the first is a forward propagation, in which the signals from a training set sample are propagated layer by layer, and during the second phase, the errors are propagated in a recursive manner while the weights are adjusted through some adjustment rule [12], [13]. The steepest descent algorithm is commonly used to tune the MLP, in which the derivatives are calculated via the backpropagation method [15].

Remark 2: Extreme Learning Machines (ELM). As well as MLP networks, ELM is a feedforward model with a single intermediate layer. However, they are unorganized machines (UM) whose intermediate neurons have weights that are chosen in a random and independent way and, the training process sums up in finding the best set of weights of the output layer. In UMs, the training process is simpler and computationally efficient because the neurons of the single intermediate layer remain untrained without losing performance [24]. Therefore, the adjustment process is solely involves finding the best set of neuron weights in the output layer, which is, a linear combiner [14], [15]. To realize this task, Huang et al. [24] suggest the use of the Moore-Penrose generalized inverse operation as it simultaneously minimizes the norm of the output weight vector and the mean square errors between the network output and the desired signal.

2.4 Computational Details

The computational step involved the six input variables, and the desired signal (target) was morbidity (number of hospital admissions due to cardiovascular diseases). The performances were evaluated considering all inputs at the same time. It is important to consider that the impact of air pollution on human health in terms of hospital attendance can happen a few days after exposure. Therefore, this analysis was performed from zero to seven days after exposure (lag days), as it is common to investigate in epidemiological studies [25]. The two neural models previously described were applied: MLP and ELM.

The data were divided into three sets:

Training: from 09/01/2018 to 06/19/2019 (349 samples);

Validation: from 06/22/2019 to 10/06/2019 (100 samples);

Test: from 10/07/2019 to 01/24/2020 (100 samples).

During the test, 30 simulations were performed for each series. The number of neurons was defined empirically by previous tests, initiating with three, then five and after that with an increase by increments of five until reaching 200 units. All the ANN present only one hidden layer, with the linear identity function in the output neurons as an activation function, and the hyperbolic tangent in the hidden layer. In addition, the weights were generated randomly in the range [-1; 1] as well as data were normalized in the same interval The adopted performance metrics were the Mean Square Error (MSE), the Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE) [26]. For reference, the model is more accurate as lower are the MSE and MAPE measures.

3 Case study

During this period, the daily number of hospital admissions varied from 3 to 29 for cardiovascular diseases, $PM_{2.5}$ concentrations ranged from 0.17 – 35 µg.m^{-3}, mean temperature from 13 to 35ºC, relative humidity from 54 to 92%, and daily precipitation summed up to 110 mm.

The city presents $PM_{2.5}$ concentrations above both (annual mean of 5µg m^{-3} and 15 µg m^{-3} 24-h mean) of the Air Quality Guidelines established by the World Health Organization. In the study period, September was the month with the highest morbidity (15 ± 6), while the minimum occurred in December (11 ± 4). In Fig. 1, the behavior of the morbidity database during the studied period is shown. It is possible to observe the seasonal behavior of the morbidity data, as well as the influence of the temporal variable "day of the week", in which the number of hospital admissions increases from Monday (10 ± 3) to Sunday (16 ± 5).

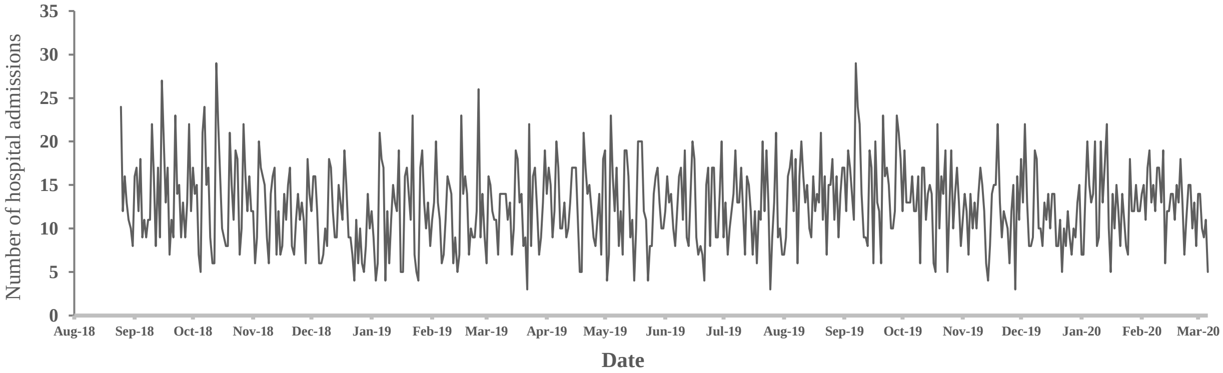

Figure 1. Number of hospital admissions for cardiovascular diseases during the period of study

The models' performance is summarized in Tab. 1. The achieved values are from the best results among 30 independent simulations and the best performance is highlighted in bold. Considering the conflicting results based on distinct error metrics to evaluate the performance, the best configuration was selected prioritizing the MSE value in the validation set since it penalizes higher errors [11], [13]–[15]. Therefore, the Multilayer Perceptron (MLP) architecture achieved the best overall results for morbidity, with a Mean Square Error (MSE) of 16.9.

Table 1. Computational results for cardiovascular diseases using ANN. "Lag" indicates the corresponding lag days from exposure, "ANN" being the acronym for the Artificial Neural Network model tested, "NN" the number of neurons in the neural models' hidden layer

Lag	ANN	NN	MSE	MAE	MAPE
0	ELM	50	36.2	4.86	38.8
	MLP	20	17.9	3.30	34.5
1	ELM	20	20.7	3.54	31.3
	MLP	20	17.1	3.20	34.5
2	ELM	40	30.2	4.43	41.6
	MLP	20	17.9	3.31	36.0
3	ELM	50	28.4	4.29	39.5
	MLP	**20**	**16.9**	**3.26**	**35.6**
4	ELM	40	26.4	4.18	35.9
	MLP	20	17.9	3.39	33.8
5	ELM	50	21.5	3.71	33.3
	MLP	20	17.0	3.30	33.3
6	ELM	30	26.3	4.06	33.0
	MLP	20	17.9	3.37	33.5
7	ELM	20	32.5	4.62	36.8
	MLP	20	18.4	3.40	36.8

As shown in Fig. 2, the neural model achieving the best results also presents the smallest dispersion in the boxplot of the MSE, demonstrating that the MLP model is the best suitable for the prediction of cardiovascular diseases in this case study

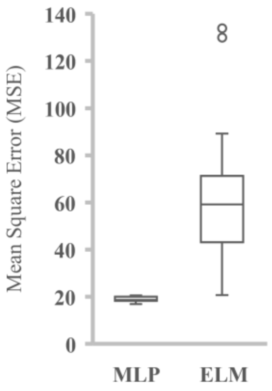

Figure 2. Boxplot of MSE of the best scenario for MLP (lag 3) and ELM (lag 1)

Friedman's test with a 95% confidence interval was applied to the MSE of the test sets and a p-value below 0.01 ($4\ 10^{-8}$) was achieved, which indicates a significant difference when changing the neural model. In addition, to assure that using ANN is the best choice, we applied a Generalized Linear Model (GLM) with Poisson Regression. As expected, the achieved errors were higher than for either neural model (MSE = 158, MAE = 110.5, and MAPE = 87). This better performance of ANN was also observed in Campinas and São Paulo by Araújo et al. [15], when exploring respiratory health effects of ambient PM_{10} exposure, and by Kassomenos et al. [27] in Athens, when studying cardiorespiratory hospital admissions due to air pollution exposure.

The MLP neural model achieved better results (i.e., lowest MSE) than ELM for all lags. It corroborates with the results achieved by Polezer et al. [10], Araújo et al. [15], and Tadano et al. [28] in other air pollution epidemiological studies exploring respiratory health effects of air pollution exposure in Curitiba [10], Campinas and São Paulo [15; 28]. However, it is important to highlight that the best results may be given by different models as it depends on the dataset behavior and input variables considered, as it was observed by Kachba et al. [11] and Tadano et al. [28].

4 Conclusions

The levels of air pollution given by the observed $PM_{2.5}$ concentrations during the study period in Joinville/Santa Catarina are of concern to public health since they exceeded the recommended levels. Moreover, the lack of information about air quality due to financial resources is a reality in most developing countries, like Brazil, hindering the use of traditional statistical regressions to estimate the impact of air quality on human health.

To the best of our knowledge, this is the first study on predicting hospital admissions for cardiovascular diseases due to air pollution using Artificial Neural Networks (ANN). Even though this is a preliminary investigation, ANN performed better than traditional statistical regressions considering the database behavior. Therefore, this study points out the major contribution that ANN can bring to epidemiological studies as they make it possible to assess impacts even when traditional models do not fit well.

In this study, we employed ANN using unpreprocessed raw data and since seasonal and trend patterns can have significant impact on various forecasting methods, data processing may improve even more the forecasting performance. As with most known methods, the main difficulty of forecasting extreme events of the number of hospital admissions remains. These challenges will be explored in future studies.

Weather patterns have great impacts on pollutants concentration, leading to similarities or differences in their temporal and spatial patterns over broad regions. Therefore, it is proposed that the performance of ANNs for forecasting cardiovascular health effects of air pollution exposure should be tested with datasets from different locations and, by exploring its capacity to deal with limited data, even grouped by seasons (when a sufficiently large dataset is available).

Acknowledgements. The authors thank the Federal University of Technology - Parana, and Brazilian agencies: Coordination for the Improvement of Higher Education Personnel - Brasil (CAPES) - Finance Code 001, Brazilian National Council for Scientific and Technological Development (CNPq) (processes number 40558/2018-5, 315298/2020-0), and Araucaria Foundation, process number 51497, for their financial support. The authors gratefully acknowledge the Santa Catarina Civil Defense for the meteorological data used in this publication

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work. and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors or has the permission of the owners to be included here.

References

[1] P. J. Landrigan *et al.*, “The Lancet Commission on pollution and health,” *The Lancet*, vol. 391, no. 10119, pp. 462–512, Feb. 2018, doi: 10.1016/S0140-6736(17)32345-0.

[2] WHO, “Billions of people still breathe unhealthy air: new WHO data,” *World Health Organization*, 2022. https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data (accessed Apr. 10, 2022).

[3] WHO, “Ambient (outdoor) air pollution,” *World Health Organization*, 2021. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed Feb. 15, 2021).

[4] IHME, “Resources | State of Global Air,” 2020. https://www.stateofglobalair.org/resources (accessed Mar. 02, 2022).

[5] L. G. Ardiles *et al.*, “Negative Binomial regression model for analysis of the relationship between hospitalization and air pollution,” *Atmospheric Pollution Research*, vol. 9, no. 2, pp. 333–341, Mar. 2018, doi: 10.1016/j.apr.2017.10.010.

[6] R. Feng *et al.*, “Recurrent Neural Network and random forest for analysis and accurate forecast of atmospheric pollutants: A case study in Hangzhou, China,” *Journal of Cleaner Production*, vol. 231, pp. 1005–1015, Sep. 2019, doi: 10.1016/j.jclepro.2019.05.319.

[7] Y. S. Tadano, C. M. Ugaya, and A. Teixeira, “Methodology to Assess Air Pollution Impact on Human Health Using the Generalized Linear Model with Poisson Regression,” in *Air Pollution - Monitoring, Modelling and Health*, M. Khare, Ed. InTech, 2012. doi: 10.5772/33385.

[8] M. Zhao, Y. Liu, and A. Gyilbag, “Assessment of Meteorological Variables and Air Pollution Affecting COVID-19 Cases in Urban Agglomerations: Evidence from China,” *IJERPH*, vol. 19, no. 1, p. 531, Jan. 2022, doi: 10.3390/ijerph19010531.

[9] I. da Silva, D. S. de Almeida, E. M. Hashimoto, and L. D. Martins, “Risk assessment of temperature and air pollutants on hospitalizations for mental and behavioral disorders in Curitiba, Brazil,” *Environ Health*, vol. 19, no. 1, p. 79, Dec. 2020, doi: 10.1186/s12940-020-00606-w.

[10] G. Polezer *et al.*, “Assessing the impact of PM2.5 on respiratory disease using artificial neural networks,” *Environmental Pollution*, vol. 235, pp. 394–403, Apr. 2018, doi: 10.1016/j.envpol.2017.12.111.

[11] Y. Kachba, D. M. de G. Chiroli, J. T. Belotti, T. Antonini Alves, Y. de Souza Tadano, and H. Siqueira, “Artificial Neural Networks to Estimate the Influence of Vehicular Emission Variables on Morbidity and Mortality in the Largest Metropolis in South America,” *Sustainability*, vol. 12, no. 7, p. 2621, Mar. 2020, doi: 10.3390/su12072621.

[12] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Ontario, Canada: Pearson Education, 2009.

[13] H. Siqueira, L. Boccato, R. Attux, and C. Lyra, “UNORGANIZED MACHINES FOR SEASONAL STREAMFLOW SERIES FORECASTING,” *Int. J. Neur. Syst.*, vol. 24, no. 03, p. 1430009, May 2014, doi: 10.1142/S0129065714300095.

[14] H. Siqueira, L. Boccato, I. Luna, R. Attux, and C. Lyra, “Performance analysis of unorganized machines in streamflow forecasting of Brazilian plants,” *Applied Soft Computing*, vol. 68, pp. 494–506, Jul. 2018, doi: 10.1016/j.asoc.2018.04.007.

[15] L. N. Araujo, J. T. Belotti, T. A. Alves, Y. de S. Tadano, and H. Siqueira, “Ensemble method based on Artificial Neural Networks to estimate air pollution health risks,” *Environmental Modelling & Software*, vol. 123, p. 104567, Jan. 2020, doi: 10.1016/j.envsoft.2019.104567.

[16] IBGE, “Cidades@ Joinville, 2021.,” *Instituto Brasileiro de Geografia e Estatística.*, 2021. https://cidades.ibge.gov.br/brasil/sc/joinville/panorama (accessed Mar. 02, 2022).

[17] R. de F. Saldanha, R. R. Bastos, and C. Barcellos, “Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS),” *Cad. Saúde Pública*, vol. 35, no. 9, p. e00032419, 2019, doi: 10.1590/0102-311x00032419.

[18] MESONET, “ASOS-AWOS-METAR Data Download.” https://mesonet.agron.iastate.edu/request/download.phtml?network=BR__ASOS (accessed Mar. 02, 2022).

[19] CEMADEM, “Baixar dados,” 2021. http://www2.cemaden.gov.br/mapainterativo/download/downpluv.php (accessed Mar. 02, 2022).

[20] C. Azorin-Molina, J. A. Guijarro, T. R. McVicar, B. C. Trewin, A. J. Frost, and D. Chen, "An approach to homogenize daily peak wind gusts: An application to the Australian series," *International Journal of Climatology*, vol. 39, no. 4, pp. 2260–2277, 2019, doi: 10.1002/joc.5949.

[21] J. A. Guijarro, "climatol: Climate Tools (Series Homogenization and Derived Products). R package version 3.1.2." 2019. Accessed: Oct. 13, 2020. [Online]. Available: https://CRAN.R-project.org/package=climatol

[22] L. N. De Castro, *Fundamentals of natural computing: an overview.*, 1st ed., vol. 4. 2019.

[23] H. Siqueira and I. Luna, "Performance comparison of feedforward neural networks 588 applied to streamflow series forecasting.," *Mathematics in Engineering*, vol. 10, no. 1, 2019.

[24] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal Approximation Using Incremental Constructive Feedforward Networks With Random Hidden Nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006, doi: 10.1109/TNN.2006.875977.

[25] Y. Li, Z. Ma, C. Zheng, and Y. Shang, "Ambient temperature enhanced acute cardiovascular-respiratory mortality effects of PM2.5 in Beijing, China," *Int J Biometeorol*, vol. 59, no. 12, pp. 1761–1770, Dec. 2015, doi: 10.1007/s00484-015-0984-z.

[26] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*, Fifth edition. Hoboken, New Jersey: John Wiley & Sons, Inc, 2016.

[27] P. Kassomenos, M. Petrakis, D. Sarigiannis, A. Gotti, and S. Karakitsios, "Identifying the contribution of physical and chemical stressors to the daily number of hospital admissions implementing an artificial neural network model," *Air Qual Atmos Health*, vol. 4, no. 3–4, pp. 263–272, Dec. 2011, doi: 10.1007/s11869-011-0139-2.

[28] Y. S. Tadano *et al.*, "Dynamic model to predict the association between air quality, COVID-19 cases, and level of lockdown," *Environmental Pollution*, vol. 268, p. 115920, Jan. 2021, doi: 10.1016/j.envpol.2020.115920.