

# Predictive modeling of optimal sites for biogas plant deployment in sugarcane agroindustrial areas using geographic data and artificial intelligence

Dados geográficos e inteligência artificial na predição de locais favoráveis para a instalação de usinas de biogás da agroindústria canavieira

Données géographiques et intelligence artificielle sont utilisées pour anticiper les emplacements propices à l'installation d'usines de biogaz dans l'agroindustrie de la canne à sucre



**Marlísia D'Abadia de Pina**

Federal Institute of Education, Science, and Technology of Goiás, Goiânia, Goiás, Brasil

[pina.marlisia@academico.ifg.edu.br](mailto:pina.marlisia@academico.ifg.edu.br)



**Édipo Henrique Cremon**

Federal Institute of Education, Science, and Technology of Goiás, Goiânia, Goiás, Brasil

[edipo.cremon@ifg.edu.br](mailto:edipo.cremon@ifg.edu.br)

**Abstract** Aligned with the imperative of the UN 2030 Agenda to facilitate the widespread adoption of renewable energies, this study underscores the pertinence of agricultural biomass, particularly derived from sugarcane, as a substantive solution to Brazil's ongoing energy transition. The determination of optimal sites for the

deployment of biogas plants is inherently contingent upon geographic considerations. This research advocates for the integration of geographic data with Artificial Intelligence algorithms, referred to as GeoAI, as a robust and prospective methodology for precisely anticipating these optimal locations. In consideration of the aforementioned, this study endeavors to forecast optimal sites for the implementation of sugarcane biogas plants within the agro-industry. By leveraging geographical data encompassing physical, biotic, and anthropic facets, alongside the utilization of six distinct classification algorithms (CART, C4.5, C5.0, Random Forest, XGBoost, and GBM), performance comparison becomes paramount. The training phase specifically targeted the state of São Paulo, owing to its heightened concentration of plants, with the most efficacious model subsequently applied to the state of Goiás. The preeminent performance achieved by the Random Forest algorithm underscores its efficacy in delineating advantageous sites for the deployment of sugarcane biogas plants in Goiás. This methodological approach holds promise in streamlining decision-making processes, delineating regions conducive to biogas production from sugarcane, thereby optimizing biomass utilization and concurrently mitigating environmental impact and installation expenditures. The incorporation of GeoAI not only fosters the proliferation of renewable energies but also contributes substantively to climate change mitigation, thereby catalyzing the broader global energy transition.

**Keywords:** Machine learning; Biomass; Renewable energy; Geographic information system.

**Resumo** A Agenda 2030 da ONU visa promover o aumento das energias renováveis em todo o mundo, e a biomassa agrícola, em particular, a partir da cana-de-açúcar, é uma solução relevante para essa transição energética no Brasil. A localização geográfica desempenha um papel crucial na determinação do local ideal para a instalação de usinas de biogás, e a combinação de dados geográficos e algoritmos de Inteligência Artificial, conhecida como GeoIA, oferece uma abordagem promissora para prever esses locais ideais. Nesse sentido, este estudo teve por objetivo prever locais favoráveis para a instalação de usinas de biogás proveniente da cana-de-açúcar da agroindústria, utilizando dados geográficos associados a aspectos físicos, bióticos e antrópicos, além de seis tipos de algoritmos de classificação (CART, C4.5, C5.0, Random Forest, XGBoost e GBM) para comparar seus desempenhos. O treinamento foi feito para o estado de São Paulo, devido ao número maior de usinas contidas na unidade federativa, e o modelo de melhor desempenho foi aplicado para o estado de Goiás. O algoritmo Random Forest obteve o melhor desempenho e permitiu identificar locais favoráveis para a instalação de usinas de biogás de cana-de-açúcar em Goiás. Essa abordagem pode facilitar a tomada de decisões ao identificar regiões propícias para a produção de biogás a partir de cana-de-açúcar, otimizando o

uso de biomassa, reduzindo o impacto ambiental e os custos de instalação. O uso de GeolA contribui para a expansão das energias renováveis e a mitigação das mudanças climáticas, promovendo a transição energética global.

**Palavras-Chave:** Aprendizado de máquina; Biomassa; Energia renovável; Sistema de informações geográficas.

**Resume** L'Agenda 2030 des Nations Unies vise à promouvoir l'essor des énergies renouvelables à l'échelle mondiale. La biomasse agricole, notamment celle issue de la canne à sucre, apparaît comme une solution pertinente pour la transition énergétique au Brésil. La localisation géographique revêt un rôle crucial dans la détermination des sites optimaux pour l'implantation de centrales de biogaz. La combinaison de données géographiques et d'algorithmes d'Intelligence Artificielle, regroupés sous le terme de GeolA, offre une approche prometteuse pour anticiper ces emplacements privilégiés. Dans cette perspective, l'objectif de cette étude était de prédire les sites favorables à l'installation de centrales de biogaz issues de la canne à sucre dans l'agro-industrie. Ceci a été réalisé en utilisant des données géographiques associées à des aspects physiques, biotiques et anthropiques, ainsi que six types d'algorithmes de classification (CART, C4.5, C5.0, Random Forest, XGBoost et GBM) afin de comparer leurs performances. L'entraînement a été effectué dans l'État de São Paulo, en raison du nombre prépondérant d'installations dans cette unité fédérative, et le modèle présentant les meilleures performances a été appliqué à l'État de Goiás. L'algorithme Random Forest a démontré les meilleures performances, permettant ainsi d'identifier les sites favorables à l'implantation de centrales de biogaz issues de la canne à sucre à Goiás. Cette approche facilite la prise de décision en identifiant les régions propices à la production de biogaz à partir de la canne à sucre, optimisant ainsi l'utilisation de la biomasse tout en réduisant l'impact environnemental et les coûts d'installation. L'utilisation de GeolA contribue à l'expansion des énergies renouvelables et à l'atténuation du changement climatique, favorisant ainsi la transition énergétique à l'échelle mondiale.

**Mots clés:** Apprentissage automatique; Biomasse; Énergie renouvelable; Système d'information géographique.

## Introduction

The pursuit of sustainability has emerged as a paramount concern within the global discourse, encapsulating the quest for equilibrium between the utilization of natural resources and the preservation of ecological integrity. At the forefront of this endeavor is the 2030 Agenda, comprising 17 Sustainable Development Goals (SDGs), initiated by the United Nations (UN) in 2015. Central to these goals is SDG 7, which delineates objectives aimed at enhancing the accessibility and cleanliness of energy sources worldwide. Specifically, SDG 7 seeks to substantially augment the proportion of renewable energy within the global energy portfolio, accelerate the pace of energy efficiency enhancements, facilitate access to clean energy research, technology, and investment, and modernize energy infrastructure in developing nations (UN, 2015).

Within this framework, the energy sector is witnessing a transformative shift, characterized by the emergence of biofuel production through clean technologies aimed at supplanting fossil fuel dependency and substantially curtailing greenhouse gas (GHG) emissions (Romero et al., 2023). Furthermore, the finite nature of non-renewable energy reservoirs underscores the urgency of transitioning towards sustainable alternatives. Research indicates that at current consumption rates and projected energy demands, reserves of oil, coal, and gas are anticipated to be depleted within approximately 35, 107, and 37 years respectively (Ioannou et al., 2018). Consequently, there is a growing imperative to embrace renewable energy sources that not only mitigate climate impacts but also derive from natural resources capable of self-renewal within a timeframe compatible with human existence, ensuring their availability without depletion.

The burgeoning market share of renewable energy is propelled by the utilization of bioproducts such as bioethanol, biodiesel, and biogas, derived from agricultural biomass (Romero et al., 2023). One notable

reservoir of waste ripe for conversion into biogas is the sugarcane agroindustry. The gross production value (GPV) of sugarcane surged to 80 billion reais in 2022 (CNA, 2023), constituting nearly 1% of Brazil's GDP, which stood at an estimated 9.9 trillion reais (IBGE, 2023). With such substantial production volumes, this sector boasts immense potential for organic waste, rendering it a lucrative reservoir of raw materials.

The utilization of biogas as an energy source stands as a promising alternative to conventional energy reservoirs, heralding a multifaceted array of benefits across environmental, economic, and social domains. Serving as a versatile energy matrix, biogas fosters sustainable development by virtue of its capacity to harness renewable waste sources for production.

In addition to its pivotal role in combatting climate change, biogas serves as a cornerstone in decentralized energy dissemination. Its adaptable nature allows for seamless integration across varying production scales and geographical contexts, accommodating diverse energy applications (Jende et al., 2016). Moreover, its utilization is instrumental in fostering the principles of the circular economy. Beyond merely providing renewable energy, biogas epitomizes a regenerative and restorative process, effectively curtailing waste generation through the reuse and recycling of raw materials and its own byproducts (Romero et al., 2023).

Owing to the multitude of factors necessitating consideration—including proximity to consumer markets, raw material availability, infrastructure adequacy, and regulatory frameworks—the identification of optimal sites for biogas plant placement remains a complex undertaking (Ioannou et al., 2018). In this context, the amalgamation of artificial intelligence (AI) methodologies, specifically machine learning algorithms, with geographic data emerges as a potent tool for enhancing the efficacy of biogas utilization planning, thereby fostering sustainable development.

The fusion of artificial intelligence (AI) methodologies with geographic data engenders GeoAI, constituting a subfield within spatial data science renowned for its capacity to facilitate the extraction of geographic insights with enhanced intelligence (Janowicz et al., 2020). Through its adept integration, GeoAI holds the potential to optimize biogas utilization by discerning favorable locations for biogas plant deployment, thereby enabling more comprehensive and precise analytical assessments. This synergy enables the amalgamation of environmental, social, and economic parameters with spatial data, culminating in the development of accurate models conducive to advancing sustainable development objectives (Romero et al., 2023).

Algorithms play a pivotal role within GeoAI, particularly in forecasting optimal sites for biogas plant deployment. Their inherent capability to process multiple variables concurrently and iteratively refine predictions through continuous learning from available datasets culminates in the generation of robust and dependable models essential for managerial decision-making processes.

The research objective is to delve into bioenergy alternatives with a specific emphasis on mitigating environmental footprints. Biogas stands out as a viable avenue for clean energy production, thereby aligning with sustainable development imperatives. The study seeks to harness the power of machine learning algorithms in tandem with geographical data to prognosticate optimal sites conducive to the installation of biogas plants within the sugarcane agro-industry.

## Methodology

Eight variables (see Table 1) are defined in this study to predict suitable locations for the installation of biogas plants in the sugarcane agro-industry. The variables were selected based on a literature review.

**Table 1:** Presents the variables used to predict suitable locations for installing a biogas plant that uses sugarcane waste as a source.

Variables/Authors	A	B	C	D	E	F	G	H	I	J
<b>Distance from productive area (agriculture and forestry)</b>				X		X			X	
<b>Distance from urban areas</b>		X	X	X	X	X	X	X	X	X
<b>Distance from sugar cane areas</b>	X			X	X	X	X	X	X	X
<b>Distance from roads and railroads</b>	X	X	X	X	X	X	X	X	X	X
<b>Distance from native vegetation</b>	X	X	X	X	X	X	X	X	X	X
<b>Distance from hydrography</b>		X	X	X		X	X	X	X	X
<b>Distance from power lines</b>	X	X	X	X	X	X		X	X	X
<b>Slope</b>			X	X		X		X	X	X

The data is derived from various bibliographic sources, including Dagnall, Hill and Pegg (2000), Sliz-Szkliniarz and Vogt (2012), Sultana and Kumar (2012), Silva, Alçada-Almeida and Dias (2014), Franco et al. (2015), Ioannou et al. (2018), Laasasenaho et al. (2019), Lozano-García et al. (2020), Yalcinkaya (2020), and Zhao et al. (2022). (2015), Ioannou et al. (2018), Laasasenaho et al. (2019), Lozano-García et al. (2020), Yalcinkaya (2020), and Zhao et al. (2022). (2015), Ioannou et al. (2018), Laasasenaho et al. (2019), Lozano-García et al. (2020), Yalcinkaya (2020), and Zhao et al. (2022).

Source: Own authorship

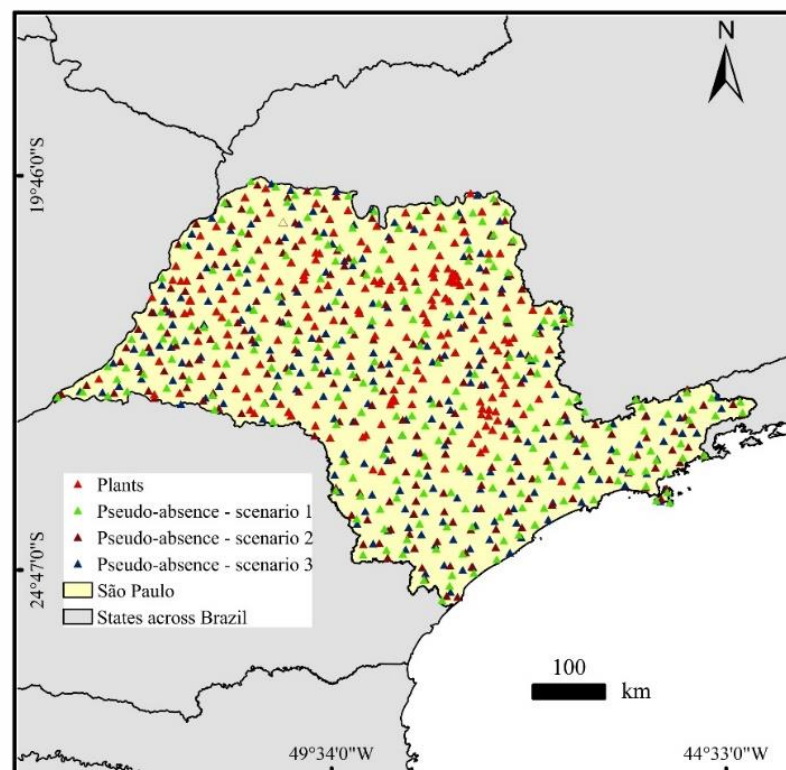
In order to train an artificial intelligence model, data from the state of São Paulo in southeastern Brazil was employed. This state exhibits a notably higher concentration of sugarcane power plants, totaling 214 units, in comparison to other states across Brazil.

The geospatial coordinates of the plant were integrated into a vector layer of points through the utilization of QGIS software (version 3.28). This methodological step was undertaken to facilitate the creation of a file capable of consolidating the pixel values derived from the matrix layers

utilized in predictive modeling, as delineated in Table 1. Additionally, 215 pseudo-absence points were systematically generated within regions void of power plants. The term "pseudo-absence points" is frequently invoked in modeling endeavors where spatial data concerning the presence of a phenomenon or target is available, yet corresponding absence data is absent, a practice acknowledged in contemporary literature (Zurell et al., 2020).

For comprehensive examination of pseudo-absence point distribution, three distinct vector layers were generated: scenario 1, scenario 2, and scenario 3 (refer to Figure 1).

**Figure 1:** Sugarcane power plants and pseudo-absence points considered in the three scenarios.



Source: Own authorship.

For the geographic database of São Paulo state, distances to key features such as productive areas (agriculture and forestry), sugarcane



fields, urban zones, and native vegetation were computed utilizing MapBiomas collection 8 data, specifically the 2020 Annual Series of Land Cover and Land Use Maps of Brazil (Souza et al., 2020). To determine distances concerning hydrography, energy infrastructure, road networks, and railways, datasets from the Department of Water and Electricity (DAEE) (2008) and OpenStreetMap (OSM, 2023) were utilized. Euclidean distances were then computed in raster format with a 30-meter pixel resolution using a Geographic Information System (GIS) environment. Slope data was extracted from the Copernicus Digital Elevation Model, employing the same 30-meter pixel resolution. All datasets were projected in metric coordinates using the Albers equivalent conic projection.

Six algorithm types were employed in this study: CART (Classification and Regression Trees), C4.5, C5.0, Random Forest, XGBoost, and GBM (Gradient Boosting Machine). The model underwent training and calibration utilizing the k-fold cross-validation methodology with  $k=10$ , utilizing 70% of the dataset for this purpose (Kuhn & Johnson, 2013).

To ascertain the reliability of the classification modeling results, independent samples from the training set, constituting 30% of the total dataset, were utilized. These samples were evaluated employing performance metrics, notably the ROC (Receiver Operating Characteristic) curve and the AUC (Area Under the Curve) value (Kuhn & Johnson, 2013).

The ROC curve and AUC metric serve as standard tools for evaluating the performance of binary classification models. The ROC curve illustrates the true positive rate (sensitivity) versus the false positive rate (specificity) for the classification task (Kuhn & Johnson, 2013; Boehmke & Greenwell, 2019). Meanwhile, the area under the ROC curve (AUC) quantifies the model's ability to discriminate between positive and negative classes. AUC values range between 0 and 1, where 0 signifies a model with no predictive capability and 1 denotes a perfect classifier. Higher AUC values correspond

to superior predictive performance in binary classification tasks. The analysis in this study was conducted using the R language and the RStudio interface (Kuhn & Johnson, 2013; Boehmke & Greenwell, 2019).

To assess the significance of variables in predicting appropriate locations for sugarcane-based biogas plant installation, a function was utilized to quantify the weights assigned to each variable during algorithm development in modelling (Kuhn; Johnson, 2013). These coefficients are normalized on a scale of 0 to 100 for each scenario used, with higher values indicating greater importance of the variables in the modelling process (Kuhn; Johnson, 2013).

The algorithms employed in this study facilitated the modeling of the raster dataset, enabling the prediction of the most and least suitable locations for sugarcane biogas plant installation based on the probability of occurrence. The pixel values within the dataset spanned from 0 to 1, where 0 denoted the least favorable locations and 1 indicated the most favorable sites. To refine the analysis, a mask was applied to urban areas and native vegetation within the modeling raster, integrating pertinent land use and land cover data.

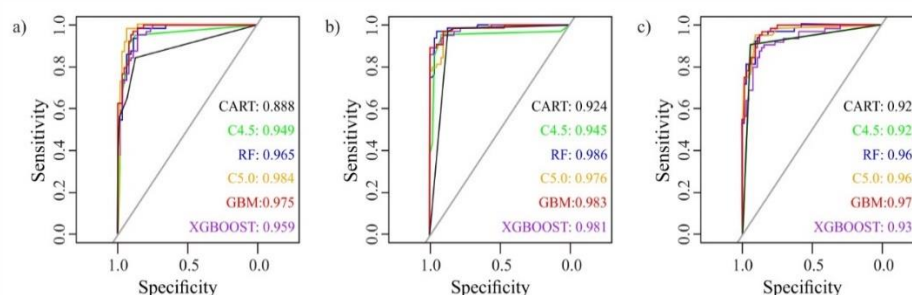
For simulation purposes, the outcomes of the best-performing model were extrapolated to the state of Goiás. This selection was informed by Goiás' status as the second-largest sugarcane producer in Brazil, trailing only São Paulo (IBGE, 2020). While the cartographic base remained consistent, minor adjustments were made to the hydrographic delineations to align with the continuous cartographic framework specific to Goiás (IBGE, 2022). To streamline computational processes during model prediction, the geographic matrix data for Goiás was resampled to a 90-meter pixel resolution, reflecting its larger geographical extent compared to São Paulo. Furthermore, the dataset was transformed into metric coordinates utilizing the Albers conic projection for spatial consistency.

Additionally, an investigation was undertaken regarding the geographical distribution of biomass thermoelectric plants operational in Goiás, utilizing sugarcane as a primary biomass source (EPE, 2023). This selection is rationalized by Goiás' current status of hosting solely one biogas plant, utilizing sugarcane waste, situated in the municipality of Goianésia. The prevalent utilization of sugarcane biomass for both electricity and biogas production underscores its significance within the region. Moreover, this validation endeavor is feasible due to the distinction between biomass power plants, which predominantly combust sugarcane bagasse, and biogas production processes, which involve filter cake and vinasse utilization—by-products not employed by biomass power plants. Consequently, an analysis of biomass plant locations enables the identification of suitable areas for biogas plant implementation, thereby optimizing the utilization of resources within the sugarcane agro-industry in Goiás.

## Results

The evaluation of algorithm performance across three distinct scenarios, as determined by the AUC metric, yielded noteworthy findings (refer to Figure 2). Each scenario underwent separate assessment to discern the algorithm's optimal performance within its unique context.

**Figure 2:** illustrates the comparison of machine learning algorithm performance (including CART, C4.5, Random Forest, C5.0, GBM, and XGBoost) across three distinct scenarios (a) scenario 1; b) scenario 2; and c) scenario 3) based on the ROC curve analysis.



Source: Own authorship.

In scenario 1 (Figure 2a), the C5.0 algorithm achieved the highest AUC of 0.984, followed closely by Random Forest with an AUC of 0.965. GBM and XGBoost had slightly lower AUCs of 0.975 and 0.959, respectively. CART and C4.5 had the lowest AUCs in this scenario, with values of 0.888 and 0.949, respectively.

Scenario 2 (Figure 2b) demonstrated that the Random Forest algorithm had the best performance, achieving an AUC of 0.986, which was the highest among all the algorithms in all scenarios. C5.0 also achieved a high AUC of 0.976, followed by GBM with 0.983. XGBoost and C4.5 also showed solid performance with AUCs of 0.981 and 0.945 respectively. CART had the lowest AUC in this scenario, with a value of 0.924.

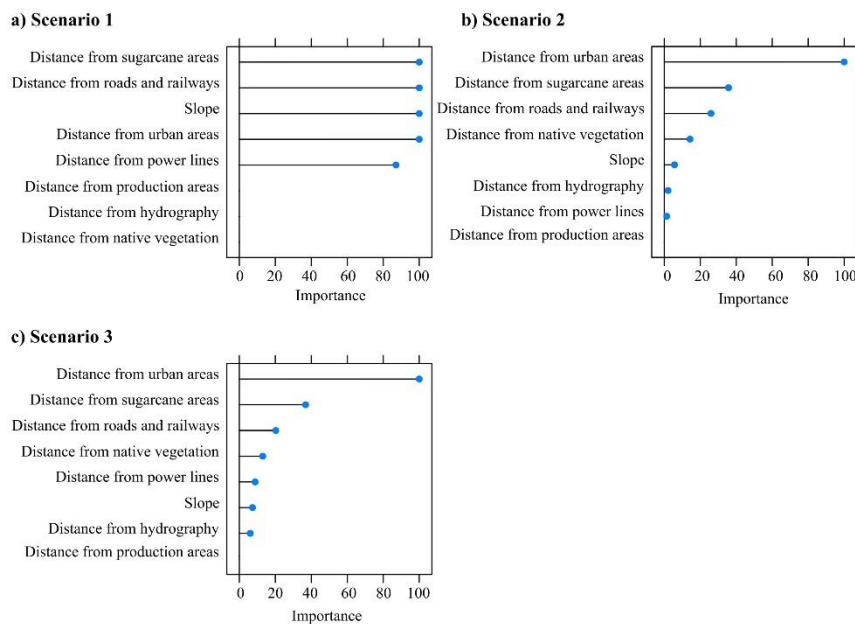
In Scenario 3 (Figure 2c), the GBM algorithm emerged as the most proficient, boasting an AUC score of 0.971. Its notable capacity to effectively discriminate between positive and negative classes renders it the optimal algorithm for this particular context. Meanwhile, the C5.0 algorithm and Random Forest both attained AUCs of 0.964 and 0.965, respectively, showcasing comparable and competitive performance. Conversely, XGBoost achieved a slightly lower AUC of 0.932 compared to the aforementioned algorithms. CART and C4.5 exhibited the lowest AUCs in this scenario, each recording a value of 0.925.

Based on the AUC results, it can be inferred that the C5.0 algorithm exhibits superior effectiveness in scenario 1, while Random Forest demonstrates the highest efficacy in scenario 2. Conversely, GBM outperforms other algorithms in scenario 3 (refer to Figure 2). This assessment underscores the significance of algorithm selection

tailored to the specific context and objectives of the application. Moreover, it emphasizes the importance of considering the scenario in which the analysis is conducted to ensure optimal performance.

In terms of variable importance depicted in Figure 3 across the three scenarios, 'Distance from Urban Area', 'Distance from Areas with Sugar Cane', and 'Distance from Roads and Railroads' emerge as the most significant factors. Conversely, there is comparatively less evidence supporting the significance of 'Distance from Productive Area'. In scenario 1, 'Declivity' and 'Distance from Power Stations' exhibit greater importance compared to scenarios 2 and 3. Conversely, 'Distance from Native Vegetation' holds equal importance in scenarios 2 and 3, but negligible importance in scenario 1. Notably, 'Distance from Hydrography' is deemed unimportant across all three scenarios (refer to Figure 3).

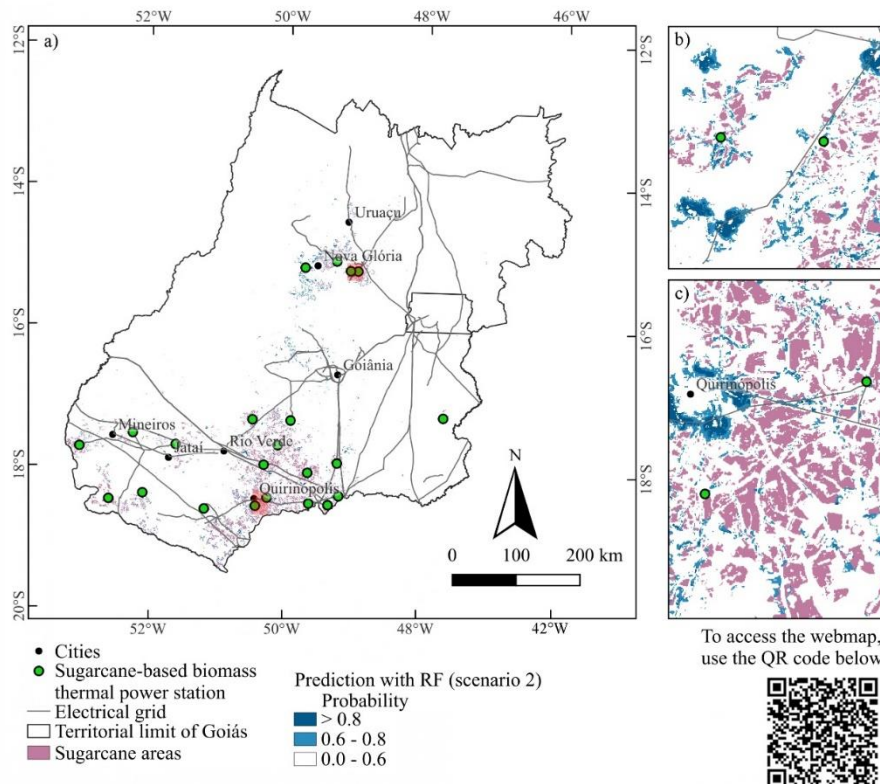
**Figure 3:** shows the importance of variables in predicting biogas plants derived from sugarcane waste in three scenarios using the C5.0, Random Forest, and GBM algorithms.



Source: Own authorship.

In scenario 2, Figure 4 displays the Random Forest algorithm's forecast of appropriate sites for sugarcane-derived biogas plant installation in Goiás state. The legend grades values from 0 (white) to 1 (blue), with pink shades indicating sugarcane-growing regions. The optimal locations for placing these mills are near sugarcane growing areas, transportation routes, and urban areas. This indicates that proximity to transportation infrastructure and the availability of sugarcane waste are the most important factors in determining the appropriate location for sugarcane-derived biogas plants.

**Figure 4:** shows the cartographic product of the prediction for the state of Goiás. The region of Nova Glória and Goianésia is highlighted in (b), where the sugarcane-derived biogas plant in operation in Goiás is represented by a green dot on the left. The region of Quirinópolis is shown in (c).



Source: Own authorship.

During the assessment of the predictive model's performance in identifying suitable sites for biogas plants derived from sugarcane waste in Goiás, it was noted that out of the 23 thermoelectric plants in Goiás utilizing sugarcane biomass, 20 exhibited probability values exceeding 0.50. This observation suggests that the predictive model, initially trained with data from São Paulo state and subsequently applied to Goiás, effectively discerned areas conducive to biogas plant installation with a high level of replicability.

Upon scrutinizing the prediction model for Scenario 2, utilizing the Random Forest algorithm, it was discerned that 94 out of 246 municipalities in Goiás state exhibit a probability exceeding 80% for favorable locations suited for the installation of sugarcane-derived biogas plants. Predominantly, the most suitable cities cluster within the southern mesoregion of Goiás. Noteworthy municipalities include Quirinópolis, Itumbiara, Cachoeira Dourada, Rio Verde, Santa Helena de Goiás, and Jataí. Additionally, within the central mesoregion of Goiás, municipalities such as Nova Glória, Goianésia, São Luiz do Norte, Itapaci, and Ipiranga de Goiás stand out as notable locations.

## Discussions

In machine learning modeling contexts, it's imperative to acknowledge that the strategic arrangement of pseudo-absence points can significantly impact algorithm performance. The positioning of pseudo-absence points across different scenarios exerted discernible effects on the performance of various algorithms. Notably, the C5.0, Random Forest, and GBM algorithms showcased superior outcomes across diverse spatial configurations.

Ensemble-based algorithms, including Random Forest, C5.0, GBM, and XGBoost, exhibited superior performance compared to

individual algorithms such as CART and C4.5. These ensemble methods amalgamate predictions from multiple models, leading to enhanced robustness and accuracy in performance (Mesri, Tahseen, & Ogla, 2021).

Upon reviewing the AUC metric results provided, it becomes apparent that Random Forest consistently attained the highest AUC scores across the board. Furthermore, GBM also demonstrated competitive AUC values, while C5.0 maintained a steady performance. These ensemble algorithms are celebrated for their capability to mitigate overfitting, effectively manage noisy data, and furnish stable predictions. Consequently, they represent a dependable choice across various machine learning applications (Mesri; Tahseen; Ogla, 2021).

Accurate data concerning transportation distance holds significant value during the planning phase, enabling precise calculations of CO<sub>2</sub> emissions, transportation expenses, and investment outlays (Höhn et al., 2014). Particularly, biogas plants exhibit optimal performance in regions characterized by higher population density when employed for electricity and heat generation purposes (Höhn et al., 2014).

The most important variables for the prediction models were "Distance from Urban Area," "Distance from Areas with Sugarcane," and "Distance from Roads and Railroads." Studies show that production and transportation costs play a critical role in determining the success or failure of bioenergy plants (Costa et al., 2020; Jayarathna et al., 2020; Latterini et al., 2020). In this sense, there is a dependence on the geographical location of the raw material, as both



the production and transportation of biomass represent a significant portion of the costs involved (Costa et al., 2020; Jayarathna et al., 2020; Latterini et al., 2020).

Goianésia, one of the cities best qualified to install biogas plants, inaugurated a cogeneration plant on September 29, 2023, utilizing vinasse from the Jalles Machado sugar-alcohol plant. It is important to note that this plant, located in Goianésia, had a probability of 0.73 in the predictive model. This value highlights the accuracy of the model in identifying suitable sites for the installation of biogas plants and confirms the viability of the strategic choice of Goianésia as the location.

Efficiently transitioning to new energy sources necessitates infrastructure development and supportive government policies. This encompasses investments in smart electricity grids, the provision of tax and regulatory incentives, and the implementation of tariff policies conducive to integrating renewable energy sources into the electricity grid.

Renewable energy stands as a pivotal element for a sustainable future, offering a clean and environmentally friendly alternative to fossil fuels. It plays a crucial role in reducing pollution, mitigating greenhouse gas emissions, and combating climate change. One of its primary advantages lies in its minimal environmental impact, being derived from natural resources with a renewable cycle that persists throughout human life without depletion. With renewable energy becoming increasingly accessible and competitive, it not only enhances energy security but also fosters local job creation.

Nevertheless, challenges persist in the adoption of renewable energy, particularly in selecting the appropriate resource and determining suitable sites for energy production. Renewable resources exhibit varying energy production potentials contingent upon their geographical location (Bravo; Casals; Pascua, 2007). The choice of resource and site can significantly influence energy output, along with associated costs, technical feasibility, and societal acceptance (Ioannou et al., 2018).

Wind energy is optimal for regions characterized by strong and consistent winds (Castro, 2004), whereas solar energy finds greater suitability in areas with high exposure to direct sunlight, as it facilitates the photovoltaic effect (Zilles et al., 2012). Hydroelectricity is most effectively harnessed in locales with high-flowing rivers (Scartazzini; Livi, 1999), while biomass emerges as a more viable option in regions abundant in agricultural and forestry waste (Ioannou et al., 2018), such as sugarcane residues from sugar and alcohol agro-industries.

Biomass offers the potential to fulfill energy requirements through direct combustion or conversion into gaseous, liquid, and/or solid fuels via thermochemical or biochemical processes. It is imperative to assess the environmental and social ramifications associated with each renewable energy source (Lozano-García et al., 2020).

The determination of resources and sites for renewable energy deployment is influenced by several factors, encompassing infrastructure availability, access to the electricity grid, installation and operational expenses, demand for renewable energy, and

governmental policies and regulations (Bravo; Casals; Pascua, 2007; Ioannou et al., 2018). Consequently, a meticulous and equitable evaluation of these factors is imperative to optimize decision-making regarding the selection of renewable energy sources and their siting. This study holds potential to contribute significantly by pinpointing factors directly affecting environmental, economic, and social variables.

A potential enhancement to the proposed methodology involves extending the application of GeoAI beyond location planning to encompass plant sizing (Yalcinkaya, 2020). Furthermore, the integration of route and location optimization studies could augment the precision of the model outcomes (Laasasenaho et al., 2019). Machine learning algorithms could also be leveraged to forecast biogas production (Li et al., 2022) and optimize plant operation post-installation (Maghami; Mutambara, 2023).

## Final Remarks

The concluding insights of this study are outlined as follows:

- The findings of this research have facilitated the identification of suitable sites for the establishment of biogas plants utilizing sugarcane in the agro-industry of Goiás.
- Among the trained algorithms, Random Forest yielded the most favorable outcome (AUC of 0.986) within scenario 2. The model, initially trained for one region, São Paulo, demonstrated applicability to another region, Goiás.

- The efficacy of the GeoAI technique was validated for this prediction, as locations featuring sugarcane biomass plants were appropriately incorporated into the algorithm's forecasts.
- Specialists may conduct on-site analyses to validate the suitability of locations suggested by machine learning models for project implementation.
- Leveraging the implemented model from this study, predictions could be generated using a similar approach for any other region across Brazil.
- The employed technique streamlines analysis and decision-making processes for assessing favorable sites for biogas production from sugarcane biomass, facilitating the identification of regions suitable for biodigester installation to optimize utilization while minimizing environmental impact and installation expenditures.

## References


BOEHMKE, Brad; GREENWELL, Brandon M. **Hands-on machine learning with R**. 1. ed. Flórida: CRC press, 2019. 484 p.

BRAVO, Javier Domínguez; CASALS, Xavier García; PASCUA, Irene Pinedo. Gis approach to the definition of capacity and generation ceilings of renewable energy technologies. **Energy policy**, Elsevier, v. 35, n. 10, p. 4879-4892, 2007.

CASTRO, Rui M.G. Introdução à energia eólica. **Lisboa: Portugal: Universidade Técnica de Lisboa**, 82 p., 2004.

CHEFAOUI, Rosa María; LOBO, Jorge Miguel. Assessing the effects of pseudo-absences on predictive distribution model performance. **Ecological modelling**, 210, 478-486, 2008.

COSTA, Fabrício Rodrigues; RIBEIRO, Carlos Antonio Alvares Soares; MARCATTI, Gustavo Eduardo; LORENZON, Alexandre Simões; TEIXEIRA, Thaisa Ribeiro; DOMINGUES, Getulio Fonseca; CASTRO, Nero Lemos



Predictive modeling of optimal sites for biogas plant deployment in sugarcane agroindustrial areas using geographic data and artificial intelligence

Marlísia D'Abadia de Pina. Édipo Henrique Cremon

Martins de; SANTOS, Alexandre Rosa dos; SOARES, Vicente Paulo; MOTA, Pedro Henrique Santos; TELLES, Lucas Arthur de Almeida; CARVALHO, José Romário de. Gis applied to location of bioenergy plants in tropical agricultural areas. **Renewable energy**, Elsevier, v. 153, p. 911-918, 2020.

CONFEDERAÇÃO DA AGRICULTURA E PECUÁRIA DO BRASIL (CNA). **Valor Bruto da Produção - VBP**. 2023. Disponível em: [www.cnabrazil.org.br](http://www.cnabrazil.org.br). Acesso em: 25 fev. 2023.

DEPARTAMENTO DE ÁGUAS E ENERGIA ELÉTRICA (DAEE). **Base cartográfica digital**, escala 1:50.000 - Projeto GISAT. São Paulo: DAEE, 2008.

DAGNALL, Steve; HILL, Jon; PEGG, David. Resource mapping and analysis of farm livestock manures — assessing the opportunities for biomass-to-energy schemes. **Bioresource technology**, Elsevier, v. 71, n. 3, p. 225-234, 2000.

EMPRESA DE PESQUISA ENERGÉTICA (EPE). **Sistema de informações geográficas do setor energético brasileiro**, 2023. Disponível em: <https://gisepeprd2.epe.gov.br/WebMapEPE/>. Acesso em: 05 nov. 2023.

FRANCO, Camilo; BOJESEN, Mikkel; HOUGAARD, Jens Leth; NIELSEN, Kurt. A fuzzy approach to a multiple criteria and geographical information system for decision support on suitable locations for biogas plants. **Applied energy**, Elsevier, v. 140, p. 304-315, 2015.


HÖHN, Jukka; LEHTONEN, Eeva; RASI, Saija; RINTALA, Jukka. A geographical information system (gis) based methodology for determination of potential biomasses and sites for biogas plants in southern finland. **Applied energy**, Elsevier, v. 113, p. 1-10, 2014.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Base cartográfica contínua do Brasil na escala de 1: 100.000**. BC100\_GODF versão 2022. Disponível em: <https://www.ibge.gov.br/geociencias/cartas-e-mapas/bases-cartograficas-continuas.html>. Acesso em: 02 set. 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Produção Agrícola Municipal. **Tabela 5457 - Área plantada ou destinada à colheita, área colhida, quantidade produzida, rendimento médio e valor da produção das lavouras temporárias e permanentes**. In: IBGE. Sidra: sistema IBGE de recuperação automática. 2020. Disponível em: <https://sidra.ibge.gov.br/tabela/5457>. Acesso em: 23 out. 2021.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Produto Interno Bruto - PIB**, 2023. Disponível em: <https://www.ibge.gov.br/explica/pib.php>. Acesso em: 25 fev. 2023.

IOANNOU, Konstantinos; TSANTOPOULOS, Georgios; ARABATZIS, Garyfallos; ANDREOPOULOU, Zacharoula; ZAFEIRIOU, Eleni. A spatial



Predictive modeling of optimal sites for biogas plant deployment in sugarcane agroindustrial areas using geographic data and artificial intelligence

Marlísia D'Abadia de Pina. Édipo Henrique Cremon

decision support system framework for the evaluation of biomass energy production locations: Case study in the regional unit of drama, Greece. **Sustainability**, v. 10, n. 2, 2018. ISSN 2071-1050.

JANOWICZ, Krzysztof; GAO, Song; MCKENZIE, Grant; HU, Yingjie; BHADURI, Budhendra. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. **International journal of geographical information science**, v. 34, n. 4, p. 625-636, 2020.

JAYARATHNA, Lasinidu; KENT, Geoff; O'HARA, Ian; HOBSON, Philip. A geographical information system based framework to identify optimal location and size of biomass energy plants using single or multiple biomass types. **Applied energy**, Elsevier, v. 275, p. 115398, 2020.

JENDE, Oliver; ROSENFELDT, Sebastian; COLTURATO, Luis Felipe de Dornfeld Braga; GOMES, Felipe Correia de Souza Pereira; PLATZER, Christoph; SERAVAL, Tathiana Almeida; HOFFMANN, Heike; CABRAL, Carolina Bayer Gomes; BURKARD, Thilo; LINNENBERG, Carsten; NAU, Daniel; PEREIRA, Amaro; MARIANI, Leidiane. Brasil. Secretaria Nacional de Saneamento Ambiental. Probiogás. **Barreiras e propostas de soluções para o mercado de biogás no Brasil / Probiogás; organizadores, Ministério das Cidades, Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH (GIZ). Brasília, DF: Ministério das Cidades, 2016. 74 p.: il. - (Desenvolvimento do mercado de biogás; 4). ISBN 978-85-7958-058-1.**


KUHN, Max; JOHNSON, Kjell. **Applied predictive modeling**. 2013. ed. New York: Springer, 2013. 600 p.

LAASASENAHO, K; LENSU, Anssi; LAUHANEN, R; RINTALA, J. Gis-data related route optimization, hierarchical clustering, location optimization, and kernel density methods are useful for promoting distributed bioenergy plant planning in rural areas. **Sustainable energy technologies and assessments**, Elsevier, v. 32, p. 47-57, 2019.

LATTERINI, Francesco; STEFANONI, Walter; SUARDI, Alessandro; ALFANO, Vincenzo; BERGONZOLI, Simone; PALMIERI, Nadia; PARI, Luigi. A gis approach to locate a small size biomass plant powered by olive pruning and to estimate supply chain costs. **Energies**, MDPI, v. 13, n. 13, p. 3385, 2020.

LI, Chao; HE, Pinjing; PENG, Wei; LÜ, Fan; DU, Rui; ZHANG, Hua. Exploring available input variables for machine learning models to predict biogas production in industrial-scale biogas plants treating food waste. **Journal of cleaner production**, Elsevier, v. 380, p. 135074, 2022.

LOZANO-GARCÍA, Diego Fabián; SANTIBAÑEZ-AGUILAR, José Ezequiel; LOZANO, Francisco J; FLORES-TLACUAHUAC, Antonio. Gis-based modeling of residual biomass availability for energy and production in Mexico.



Predictive modeling of optimal sites for biogas plant deployment in sugarcane agroindustrial areas using geographic data and artificial intelligence

Marlísia D'Abadia de Pina. Édipo Henrique Cremon

**Renewable and sustainable energy reviews**, Elsevier, v. 120, p. 109610, 2020.

MAGHAMI, Mohammad Reza; MUTAMBARA, Arthur Guseni Oliver. Challenges associated with hybrid energy systems: An artificial intelligence solution. **Energy reports**, Elsevier, v. 9, p. 924-940, 2023.

MESRI, Khitam; TAHSEEN, Israa; OGLA, Raheem. Default on a credit prediction using decision tree and ensemble learning techniques. **Journal of physics: conference series**, 2021.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS (ONU). **Transforming our world: the 2030 agenda for sustainable development**, 2015. Disponível em: <https://sdgs.un.org/2030agenda>. Acesso em: 05 nov. 2022.

OPEN STREET MAP (OSM). **OpenStreetMap database [PostgreSQL via API]**. OpenStreetMap Foundation: Cambridge, UK; 2023.


ROMERO, Cristhy Willy da Silva; MIYAZAKI, Marcelle Rose; BERNI, Mauro Donizeti; FIGUEIREDO, Gleyce Kelly Dantas Araújo; LAMPARELLI, Rubens Augusto Camargo. A spatial approach for integrating gis and fuzzy logic in multicriteria problem solving to support the definition of ideal areas for biorefinery deployment. **Journal of cleaner production**, Elsevier, p. 135886, 2023.

SCARTAZZINI, Luiz Sílvio; LIVI, Flávio Pohlmann. Potencial hidroenergético do alto rio pelotas. **Revista brasileira de recursos hídricos**, v. 4, n. 4, p. 87-95, 1999.

SILVA, Sandra; ALÇADA-ALMEIDA, Luís; DIAS, Luís C. Biogas plants site selection integrating multicriteria decision aid methods and gis techniques: A case study in a portuguese region. **Biomass and bioenergy**, Elsevier, v. 71, p. 58-68, 2014.

SLIZ-SZKLINIARZ, Beata; VOGT, Joachim. A gis-based approach for evaluating the potential of biogas production from livestock manure and crops at a regional scale: A case study for the kujawsko-pomorskie voivodeship. **Renewable and sustainable energy reviews**, Elsevier, v. 16, n. 1, p. 752-763, 2012.

SOUZA, Carlos; SHIMBO, Julia; ROSA, Marcos; PARENTE, Leandro; ALENCAR, Ane; RUDORFF, Bernardo; HASENACK, Heinrich; MATSUMOTO, Marcelo; FERREIRA, Laerte; SOUZA-FILHO, Pedro; OLIVEIRA, Sergio; ROCHA, Washington; FONSECA, Antônio; MARQUES, Camila; DINIZ, Cesar; COSTA, Diego; MONTEIRO, Dyeden; ROSA, Eduardo; VÉLEZ-MARTIN, Eduardo; WEBER, Eliseu; LENTI, Felipe; PATERNOST, Fernando; PAREYN, Frans; SIQUEIRA, João; VIERA, José; NETO, Luiz Ferreira; SARAIVA, Marciano; SALES, Marcio; SALGADO, Moises; VASCONCELOS, Rodrigo; GALANO, Soltan; MESQUITA, Vinicius; AZEVEDO, Tasso. Reconstructing



Predictive modeling of optimal sites for biogas plant deployment in sugarcane agroindustrial areas using geographic data and artificial intelligence

Marlísia D'Abadia de Pina. Édipo Henrique Cremon

three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. **Remote sensing**, v. 12, n. 17, 2020.

SULTANA, Arifa; KUMAR, Amit. Optimal siting and size of bioenergy facilities using geographic information system. **Applied energy**, Elsevier, v. 94, p. 192-201, 2012.

YALCINKAYA, Sedat. A spatial modeling approach for siting, sizing and economic assessment of centralized biogas plants in organic waste management. **Journal of cleaner production**, Elsevier, v. 255, p. 120040, 2020.

ZHAO, Bingchao; WANG, Han; HUANG, Zhihao; SUN, Qianqian. Location mapping for constructing biomass power plant using multi-criteria decision-making method. **Sustainable energy technologies and assessments**, Elsevier, v. 49, p. 101707, 2022.

ZILLES, Roberto; MACÊDO, Wilson Negrão; GALHARDO, Marcos André Barros;

OLIVEIRA, Sérgio Henrique Ferreira de. **Sistemas fotovoltaicos conectados à rede elétrica**. 1. ed. São Paulo: Oficina de textos, 2012.

ZURELL, Damaris; FRANKLIN, Janet; KÖNIG, Christian; BOUCHET, Phil Jean-François; DORMANN, Carsten; ELITH, Jane; FANDOS, Guillermo; FENG, Xiao; GUILLERA-ARROITA, Gurutzeta; GUIBAN, Antoine; LAHOZ-MONFORT, José; LEITÃO, Pedro; PARK, Daniel; PETERSON, Townsend; RAPACCIUOLO, Giovanni; SCHMATZ, Dirk; SCHRÖDER, Boris; SERRA-DIAZ, Josep; THUILLER, Wilfried; YATES, Katherine; ZIMMERMANN, Niklaus; MEROW, Cory. A standard protocol for reporting species distribution models. **Ecography**, Wiley Online Library, v. 43, n. 9, p. 1261-1277, 2020.



## Author contributions

All authors made substantial scientific and intellectual contributions to the study. The tasks of study conception, design, manuscript preparation, writing, and critical review were carried out collaboratively. The first author, Marlísia D'Abadia de Pina, was particularly responsible for the study's conception, methodology, data acquisition, interpretation and analysis, software development, validation, and data curation. The second author and corresponding author, Édipo Henrique Cremon, was responsible for theoretical and conceptual development, formal analysis, investigation, and validation. All authors contributed to the writing of the document. We also acknowledge our awareness of the General Guidelines of BGG.

**Marlísia D'Abadia de Pina** - Bachelor in Cartographic and Surveying Engineering and Master in Technology, Management, and Sustainability from the Federal Institute of Education, Science, and Technology of Goiás. Her main areas of focus include remote sensing, GIS, machine learning, and renewable energy sources.

**Édipo Henrique Cremon** - Geographer from the State University of Maringá. He holds a Master's and a Doctorate in Remote Sensing from the National Institute for Space Research (INPE), with a doctoral exchange program at the University of Exeter (United Kingdom). He is currently a Tenured Professor at the Federal Institute of Education, Science, and Technology of Goiás (IFG - Goiânia Campus) and a researcher with the Geomatics Study Group (GEO). His main areas of focus include machine learning applied to geographic data for environmental and geomorphological analysis.

Receipt date November 20, 2023

Accepted January 22, 2023

Published on May 24, 2024