

Symbolic Regression for Large-Scale Propagation Modeling in the Atlantic Forest

Pedro Victor F. Granja¹, Pedro Henrique G. Lugão², Ana Elisa Leitão A. Ferreira¹, and Diego Barreto Haddad¹

Abstract—This paper investigates the use of symbolic regression to derive analytical models for large-scale propagation in the Atlantic Forest environment using LoRa measurements. Unlike traditional empirical or machine learning approaches, symbolic regression provides concise and interpretable equations. Models are obtained using distance and spreading factor as inputs. Despite the small dataset, the derived formulas achieved a Pearson correlation coefficient above 0.98 with the measured data, indicating strong predictive capability. Sensitivity analysis reveals physically meaningful trends, such as increasing path loss with distance and improved received power with higher spreading factors. The results highlight the potential of symbolic regression as a lightweight and interpretable alternative for propagation modeling.

Keywords—Symbolic Regression, Large-scale Propagation, LoRa Networks, RSSI Modeling.

I. INTRODUCTION

Accurate modeling of small and large-scale propagation is essential for the design and optimization of wireless communication systems. In dense forest environments, signal propagation is strongly affected by absorption, scattering, and multipath effects caused by vegetation, leading to significant variability in received signal strength.

In the context of Internet of Things (IoT) and low-power wide-area networks (LPWAN), such as LoRa [1], understanding propagation in forested areas is particularly important [2]. Applications such as environmental monitoring and emergency communication depend on reliable connectivity under severe propagation conditions. Experimental studies indicate that propagation in these environments is highly sensitive to distance, vegetation, and system parameters, including the spreading factor.

Traditional propagation models are typically empirical or semi-empirical, such as log-distance models or standardized recommendations (e.g., ITU-R [3]). Although useful, they may lack flexibility and fail to capture environment-specific behavior. Conversely, machine learning methods can model complex relationships from data but often rely on black-box models with limited interpretability.

In this context, symbolic regression (SR) offers a promising alternative by discovering explicit analytical expressions

directly from data. Unlike conventional regression or neural networks, SR produces closed-form equations, enabling both prediction and interpretation, which is particularly valuable in propagation modeling [4], [5].

This paper investigates the use of symbolic regression to model large-scale propagation in the Atlantic Forest using LoRa measurements. The proposed approach considers distance and spreading factor as inputs to derive concise expressions for predicting RSSI. Due to the limited dataset, all samples are initially used for training, following common practices in propagation modeling. A controlled experiment with a held-out spreading factor is also conducted to evaluate generalization.

The main contributions of this work are: (i) the application of symbolic regression to propagation modeling in a dense forest environment using real data; (ii) the derivation of compact analytical expressions with high predictive accuracy; and (iii) an interpretability analysis revealing physically meaningful relationships between distance, spreading factor, and RSSI.

The results show that symbolic regression can bridge the gap between empirical models and black-box approaches, providing both accuracy and interpretability. Section II describes the methodology, Section III presents and discusses the results, and Section IV concludes the paper.

II. DATASET AND METHODOLOGY

A. Dataset

The dataset consists of received signal strength indicator (RSSI) measurements, in dBm, collected in a dense Atlantic Forest environment using LoRa communication links, as illustrated in Fig. 1. Measurements were obtained for multiple transmitter–receiver distances and different spreading factor (SF) configurations (see [6] for more information).

RSSI samples were collected for four spreading factors, $SF \in \{7, 9, 11, 12\}$, covering distinct LoRa link budget conditions [7]. Distances range from 0 m to 250 m, with nonuniform sampling across SFs. Measurements up to 100 m are available for all SFs, while longer distances are mainly associated with higher SFs, particularly $SF = 12$, consistent with its extended range. Although LoRa can support communication over several kilometers, the measurements are limited to shorter distances due to the strong attenuation in the dense and humid tropical forest, which significantly reduces the effective range. Overall, the dataset captures short- and medium-range propagation regimes under different configurations, providing a compact and representative basis for modeling.

The input variables are the link distance d and the spreading factor SF, both central to large-scale propagation. Distance

¹Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Brazil, ² Universidade Federal de Juiz de Fora (UFJF). Emails: pedro.granja@aluno.cefet-rj.br, pedro.lugao@ufjf.br, ana.ferreira@cefet-rj.br, diego.haddad@cefet-rj.br. This paper was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES, Finance Code 001), and in part by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

governs path loss, while the spreading factor influences link budget and receiver sensitivity [7]. In this work, SF is treated as a numerical rather than categorical variable. This choice prevents the symbolic search from fragmenting into complex piecewise branches (e.g., if-then rules). Instead, it forces the algorithm to capture the physical monotonic scaling of the link budget, where higher SF proportionally improves receiver sensitivity, into a single, compact, and continuous analytical expression.



Fig. 1

(A) AND (B) EXPERIMENTAL SCENARIO IN THE ATLANTIC FOREST. (C) PROTOTYPES USED IN THE MEASUREMENT CAMPAIGN.

B. Training Strategy

All data points were initially used to train the symbolic regression models. Although this differs from standard machine learning practice, it is consistent with propagation modeling methodologies [4], where data are often scarce and costly to obtain. This choice is justified by the small dataset size, which limits the feasibility of partitioning, and by the nature of symbolic regression, which explicitly searches for concise analytical expressions, inherently reducing model complexity and, consequently, the risk of overfitting. Moreover, the objective is not only predictive accuracy but also the extraction of interpretable relationships. To assess generalization, an additional experiment was conducted using a controlled train-test split, where all samples with $SF = 9$ were excluded from training and used exclusively for testing, evaluating the ability of the models to extrapolate across different configurations.

C. Symbolic Regression Setup

Symbolic regression was used to derive analytical expressions relating RSSI to the input variables, exploring the space of mathematical models through evolutionary strategies that balance accuracy and complexity.

Three configurations were considered: (i) an extended function set with $\log(\cdot)$, $\text{abs}(\cdot)$, and $\text{erf}(\cdot)$, enabling flexible nonlinear representations; (ii) a domain-inspired set restricted to $\log(\cdot)$ and $\text{abs}(\cdot)$, aiming at improved interpretability; and (iii) a generalization experiment using a train-test split with $SF = 9$ reserved for testing. The inclusion of $\log(\cdot)$ is motivated by

the log-distance behavior of path loss, while $\text{erf}(\cdot)$ provides a flexible mechanism to model saturation effects.

The implementation was carried out using the `PySR` library [8], with binary operators $\{+, -, \times, \div, ^\wedge\}$ and the aforementioned unary functions. Model complexity was controlled by limiting expression size to 15 and restricting nested operators. The search employed multiple populations evolving in parallel with simulated annealing and ℓ_2 loss, under a time-based stopping criterion. Empirically, the method showed robustness to hyperparameter variations, yielding consistent expressions across configurations, indicating that the models reflect intrinsic data relationships rather than parameter-specific artifacts.

III. RESULTS AND DISCUSSION

This section presents the analytical models obtained via symbolic regression and discusses their predictive performance and physical interpretability. The code is available at [link](#).

A. Models and Performance

Three symbolic regression configurations were evaluated. For comparison, a traditional log-distance Linear Regression (LR) baseline was also fitted, yielding $\text{RSSI}_{\text{LR}} = 3.15SF - 39.97 \log(d) - 64.32$. The obtained SR models are:

$$\text{Model 1: } \text{RSSI} = -88.25 \cdot \text{erf}\left(d \cdot \log(0.56^{SF} + 1.01)\right) - 35.03 \quad (1)$$

$$\text{Model 2: } \text{RSSI} = \left\lfloor \frac{151.23}{d/SF - 1.54} \right\rfloor - 133.91 \quad (2)$$

$$\text{Model 3: } \text{RSSI} = (0.9793^d \cdot 90.76) - 125.02 \quad (3)$$

Table I summarizes the performance metrics. As observed, all SR models outperform the LR baseline (which yielded an RMSE of 9.19), achieving higher correlation with the measured data and lower error margins. The SR model evaluated under the train-test split also indicates strong generalization capability.

TABLE I
PERFORMANCE METRICS FOR THE EVALUATED MODELS.

Model	RMSE	MAE	Pearson	Spearman
Linear Regression	9.19	6.88	0.966	0.853
SR (log, abs, erf)	6.59	5.40	0.983	0.846
SR (log, abs)	6.47	5.43	0.983	0.890
SR Train/Test (SF=9)	4.70	4.47	0.994	1.000

Figures 2 and 3 illustrate the agreement between measured RSSI values and the predictions provided by the symbolic regression models. In Fig. 2, Model 1 captures the overall trend of signal attenuation across different spreading factors, despite some deviations in specific regions, particularly at intermediate distances. The model is able to represent the nonlinear behavior of propagation, including saturation effects, although at the cost of increased complexity.

In contrast, Fig. 3 highlights the performance of the model obtained under the train-test split, evaluated on an unseen spreading factor ($SF = 9$). The predicted curve closely follows the measured data, demonstrating strong generalization

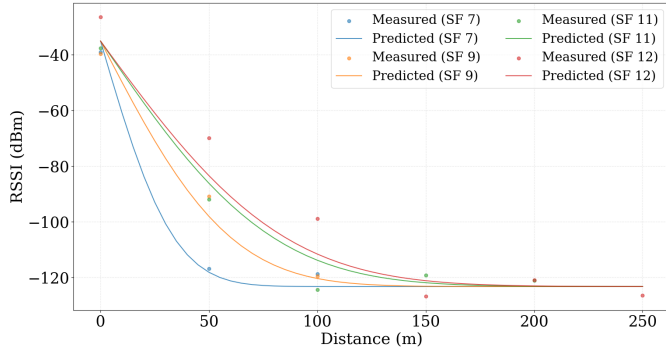


Fig. 2

MEASURED RSSI VALUES AND CORRESPONDING PREDICTIONS OBTAINED BY MODEL 1 FOR DIFFERENT SPREADING FACTORS AS A FUNCTION OF DISTANCE.

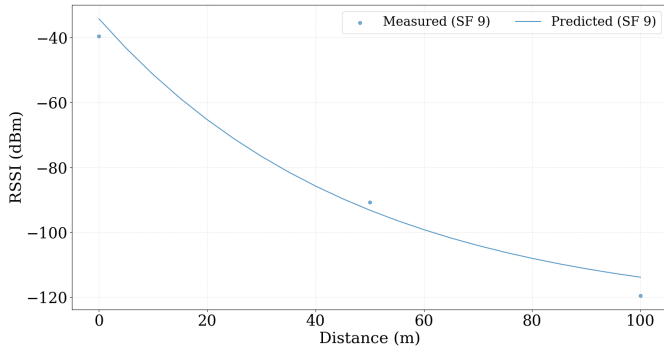


Fig. 3

MEASURED RSSI VALUES AND PREDICTED CURVE OBTAINED BY MODEL 3 ILLUSTRATING THE GENERALIZATION CAPABILITY ACROSS DISTANCE FOR AN UNSEEN SPREADING FACTOR (SF = 9).

capability even without explicit exposure to this configuration during training. This result suggests that the underlying structure captured by symbolic regression is sufficiently robust to extrapolate across system parameters, reinforcing its potential for practical propagation modeling applications.

B. Discussion

All models consistently capture a key property of large-scale propagation: RSSI decreases with increasing distance. In the first model, distance appears inside the argument of the error function, producing a smooth monotonic decay. In the second, the dependence on d arises through a rational term, where increasing distance enlarges the denominator and reduces RSSI. Interestingly, although the logarithmic function was included among the allowed primitives in this configuration, it does not appear in the final expression, suggesting that the algorithm identified an alternative functional form capable of capturing the underlying behavior without relying on the commonly used log-distance relationship. The third model encodes attenuation via an exponential decay, resembling classical propagation laws. These results indicate that symbolic regression successfully recovers fundamental physical behavior directly from

data.

The influence of the spreading factor is explicitly modeled in the first two expressions. In the first model, increasing SF reduces the term 0.5607^{SF} , altering the logarithmic component and lowering attenuation. In the second, the ratio d/SF decreases as SF increases, leading to higher RSSI. This is consistent with LoRa principles [1], where higher spreading factors improve receiver sensitivity and link budget.

A sensitivity analysis of Model 2 yields $\frac{\partial \text{RSSI}}{\partial d} < 0$ and $\frac{\partial \text{RSSI}}{\partial \text{SF}} > 0$, confirming monotonic decay with distance and improvement with spreading factor, thus reinforcing physical consistency.

The models reveal a trade-off between flexibility, interpretability, and generalization. The erf-based model offers higher nonlinear flexibility but lower interpretability. The exponential model shows good generalization under the train-test split, albeit without explicitly incorporating SF. This suggests that, under limited data and generalization constraints, symbolic regression may favor simpler and more interpretable expressions, even at the cost of neglecting relevant variables. Despite the small dataset, the method yields compact and physically meaningful expressions with strong predictive performance.

IV. CONCLUSIONS

This work demonstrated the effectiveness of symbolic regression for modeling large-scale propagation in the Atlantic Forest using LoRa measurements, yielding concise and interpretable models consistent with known physical principles. The results indicate that symbolic regression is a promising alternative to both empirical and black-box approaches, balancing predictive accuracy and interpretability. Future work includes expanding the dataset for longer distances, incorporating additional environmental variables, and evaluating robustness across different propagation scenarios.

REFERENCES

- [1] L. Alliance, "A technical overview of lora and lorawan," *White Paper*, November, 2015.
- [2] C. M. S. Figueiredo, E. F. Nakamura, A. D. Ribas, T. R. B. de Souza, and R. S. Barreto, "Assessing the communication performance of wireless sensor networks in rainforests," in *2009 2nd IFIP Wireless Days (WD)*, pp. 1–6, 2009.
- [3] ITU, "Recommendation ITU-R P.1546-6: Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 4000 MHz," tech. rep., ITU, 2019.
- [4] G. K. d. Oliveira, D. B. Haddad, G. A. Giraldi, and M. H. C. Dias, "Symbolic regression approach for analytical modeling of itu-r p.1546-6 propagation curves," *Telecommunication Systems*, vol. 88, May 2025.
- [5] G. K. de Oliveira, D. B. Haddad, G. A. Giraldi, and M. H. Costa Dias, "Analytical expression for recommendation itu-r p.1546-6 propagation curves of land paths up to 20 km using symbolic regression," in *2024 Symposium on Internet of Things (SIoT)*, vol. 1, pp. 1–5, 2024.
- [6] A. E. L. A. Ferreira, *Análise de Redes Sem-fio para Aplicações de Assistência ao Visitante de Parques Florestais*. Tese de doutorado, Universidade Federal do Rio de Janeiro (COPPE/UFRJ), Rio de Janeiro, RJ - Brasil, Março 2020. Orientador: Luís Henrique Maciel Kosmowski Costa.
- [7] A. Augustin, J. Yi, T. Clausen, and W. M. Townsley, "A study of LoRa: Long range & low power networks for the Internet of things," *Sensors*, vol. 16, no. 9, p. 1466, 2016.
- [8] A. Tonda, "Review of PySR: high-performance symbolic regression in Python and Julia," *Genetic Programming and Evolvable Machines*, vol. 26, Dec. 2024.