A Deep Learning Framework for Urban Ecoacoustic Monitoring and Management: A Synthetic-to-Real CNN-Transformer Approach

Luis J. R. Oliveira[1], Ingrid G. C. Lira[2], Maximiliano A. da S. Lopes[1]

[1]Department of Informatics, Faculty of Exact and Natural Sciences, Universidade do Estado do Rio Grande do Norte (UERN), Av. Prof. Antônio Campos, s/n, BR-110, Km 48, Costa e Silva, Mossoró, RN, Brazil

[2]Graduate Program in Computer Science, Universidade do Estado do Rio Grande do Norte (UERN) and Universidade Federal Rural do Semi-Árido (UFERSA), Av. Francisco Mota, s/n, Costa e Silva, Mossoró, RN, Brazil.

{luis.jeronimo044,ingridliragc}@gmail.com, maximilianolopes@uern.br

Abstract. *Urban acoustic monitoring generates massive volumes of continuous data, making manual soundscape inspection unfeasible at scale. This work proposes a hybrid CNN-Transformer framework for urban ecoacoustic monitoring, modeling temporal interactions among Biophony, Geophony, and Anthropophony. A Transformer encoder-only was trained on probabilities extracted by PANNs from a controlled synthetic dataset (ESC-50), eliminating dependence on weak labels. The model achieved a macro F1-Score of 0.9960 and generalized to three real-world datasets, with a Spearman correlation of 0.54 (95% CI: [0.40, 0.65], $p < 0.001$, $n = 127$) between Anthropophony predictions and measured sound levels; predictions were spatialized via Ordinary Kriging, yielding a kriging RMSE of 0.07 in Madrid. Results show that the framework captures the dynamics of biophony suppression by anthropogenic noise and provides quantitative support for urban environmental management and biodiversity conservation.*

1. Introduction

Rapid urban expansion imposes increasing pressures on ecosystems, driving habitat fragmentation and significant environmental alterations. In this context, acoustic signals emerge as sensitive indicators of these transformations, reflecting population ecology, landscape structure, and biotic integrity [Pijanowski et al. 2011; Krause 1993]. Soundscape Ecology classifies environmental sounds into three fundamental components: Biophony, Geophony, and Anthropophony. According to the Acoustic Niche Hypothesis, anthropogenic noise can induce acoustic masking and reduce the detectability of Biophony, thereby acting as a critical indicator of urban environmental degradation [Pijanowski et al. 2011; Krause 1993]. However, Passive Acoustic Monitoring (PAM) generates massive volumes of continuous data, rendering manual inspection unfeasible on an urban scale [Sugai et al. 2019]. Consequently, Machine Learning techniques have been increasingly employed to automate species identification and soundscape analysis in an efficient and scalable manner [Correia & Ventura 2025].

Approaches relying on acoustic indices or classical machine learning require manual feature extraction and exhibit severe limitations in polyphonic urban environments [Stowell 2022]. Convolutional Neural Networks (CNNs) advanced this

field by automatically extracting spectral representations [Piczak 2015; Salamon & Bello 2017], while models such as Pretrained Audio Neural Networks (PANNs) further enhanced performance through Transfer Learning [Kong et al. 2020]. Nevertheless, these approaches treat audio as independent, short-duration segments. This underestimates the long-term temporal dependencies essential for understanding biophony masking dynamics. Furthermore, static classifiers applied to PANNs representations tend to suffer from predictive fragmentation (temporal flickering) when processing continuous acoustic streams [Mesaros et al. 2021; Ronchini & Serizel 2022].

To address these gaps, this work proposes a hybrid CNN-Transformer framework to model the temporal dynamics among ecoacoustic spheres. An encoder-only Transformer is applied to the frame-wise probabilities extracted by PANNs, capturing long-term dependencies between acoustic components. The primary innovation lies in training the model exclusively with controlled synthetic data. This enables the network to learn masking effects using precise ground-truth labels, an aspect often neglected in real-world datasets reliant on weak labels [Ronchini & Serizel 2022; Quinn et al. 2022]. The resulting predictions are subsequently spatialized via Ordinary Kriging, converting probabilistic time series into continuous surfaces of urban acoustic dominance. The central hypothesis is that a model trained exclusively on synthetic data can effectively generalize to heterogeneous, real-world urban environments. Specifically, this study aims to: (i) evaluate the framework's generalization capability in real acoustic scenarios; (ii) temporally characterize the suppression of Biophony under anthropogenic pressure; and (iii) spatialize these predictions to support urban management and biodiversity conservation.

The remainder of this paper is organized as follows: Section 2 presents Related Work; Section 3 describes the Material and Methods; Section 4 discusses the Results; and Section 5 outlines the Conclusions and future perspectives.

2. Related Works

Although CNNs and PANNs have significantly advanced urban acoustic classification, structural limitations persist in the ecoacoustic domain. Quinn et al. (2022) demonstrated that the imbalance inherent in real-world data, where anthropogenic noise often masks biological signals, compromises labeling quality and hinders the distinction of patterns across ecological gradients. More critically, static classifiers applied to PANN representations suffer from predictive fragmentation when processing continuous streams. This issue is exacerbated by the reliance on weak labels, which poorly capture the boundaries of simultaneous and overlapping sound events [Mesaros et al. 2021; Ronchini & Serizel 2022], precisely the ecologically critical transition zones defined by the Acoustic Niche Hypothesis [Krause 1993].

The Self-Attention mechanism [Vaswani et al. 2017] has emerged as a robust alternative to overcome this temporal bottleneck, simultaneously weighting past and future contexts within a sequence to smooth gradual acoustic transitions. Stowell (2022) reinforces that modeling interactions between sound components in polyphonic scenarios remains an open challenge. In this context, stochastic audio synthesis [Nikolenko 2021] arises as a viable alternative to the scarcity of labeled data, enabling the creation of controlled ground-truth scenarios within transition zones. In the applied

domain, Nery et al. (2025) demonstrate the feasibility of automated systems for georeferenced urban sound analysis; Correia & Ventura (2025) establish a methodological foundation for species identification via Deep Learning with a recall exceeding 0.89; and Peralta et al. (2024) propose the semantic retrieval of soundscapes based on acoustic similarity. However, none of these works explicitly models the dynamic interaction between ecoacoustic spheres, as they treat sound events as discrete categories rather than continuous, interpolatable compositions.

This work addresses three gaps: (i) explicit temporal modeling of ecoacoustic transitions via Transformer, (ii) controlled synthetic training eliminating weak-label reliance, and (iii) spatialization via Kriging—a dimension largely unexplored in computational ecoacoustics.

3. Materials and Methods

This applied experimental research proposes a hybrid Deep Learning and geostatistical analysis framework structured into four phases (Figure 1): (i) stochastic synthesis of acoustic data, (ii) latent feature extraction via PANNs, (iii) temporal modeling using a Transformer, and (iv) spatio-temporal geostatistical visualization. The experiments were conducted on Google Colab (NVIDIA T4 GPU, 12 GB VRAM) using Python 3, alongside the Librosa, TensorFlow, Scikit-learn, and PyKrige libraries. The source code is available in a public repository for reproducibility purposes[1].

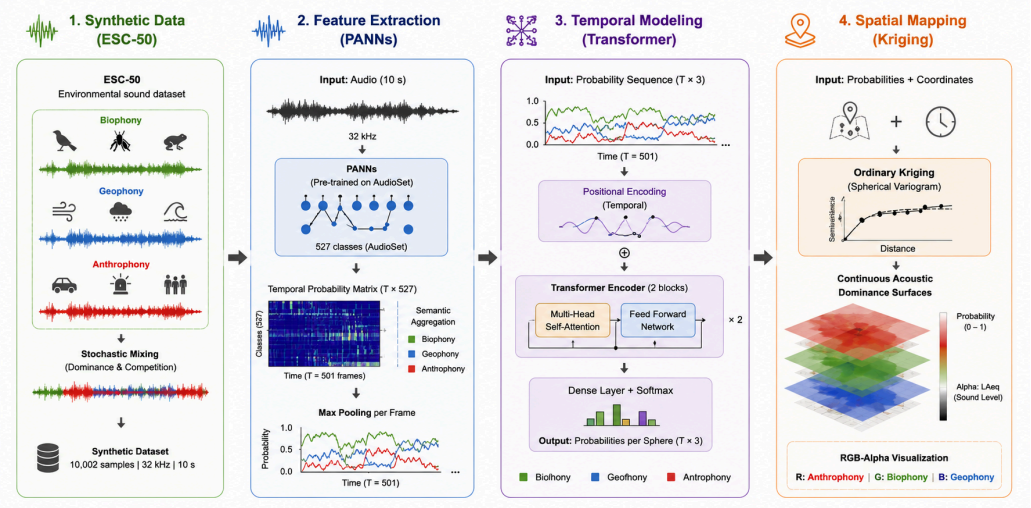


Figure 1. Integrated methodology pipeline

3.1. Synthetic Dataset Generation

A stochastic audio mixing strategy was adopted to simulate controlled ecoacoustic scenarios utilizing the ESC-50 dataset [Piczak 2015]. The classes were manually grouped into three spheres: Biophony (bird, insect, and mammal vocalizations), Geophony (wind, rain, and sea sounds), and Anthropophony (engines, sirens, and human speech). Prior to the mixing stage, the audio files were resampled to 32 kHz and normalized via Root Mean Square (RMS) to prevent amplitude biases and facilitate the

[1] Project repository: https://github.com/luisjro16/Synthetic-SoundScape

learning of spectral patterns.

Scenario generation followed a probabilistic logic under two distinct modes: (i) clear dominance scenes, where a single sphere accounted for 60% to 90% of the signal's energy, simulating isolated events; and (ii) competition scenes, where two spheres shared the dominance (ranging from 40% to 48% each), simulating acoustic transition zones or conflict scenarios. The residual energy proportions were distributed among samples from the remaining spheres to simulate realistic background noise. Scenarios featuring simultaneous co-dominance of all three spheres were excluded based on the Acoustic Niche Hypothesis [Krause 1993], as they are ecologically improbable and would introduce label noise.

Each sample was audited via PANNs (threshold: 0.4 pure, 0.25 hybrid) to preserve partial masking while discarding total source suppression—avoiding label noise that hinders generalization. Consequently, the strictness of this purity filter resulted in high rejection rates, with the stochastic mixing algorithm processing over 76,000 raw combinations to extract the clean dataset. The pipeline ultimately yielded 10,002 valid final samples, whose distribution and actual rejection rates are detailed in Table 1.

Table 1. Distribution of the Synthetic Dataset

Type	Samples	Total Duration (s)	Rejection Rate	Sound Categories (ESC-50)
Clear Dominance	5.001	25.005	66,2%	Bio (bird, insect, mammal…) / Geo (wind, rain, sea…) / Anthro (engine, siren, speech…)
Competition	5.001	25.005	91,9%	Bio–Geo, Bio–Anthro, Geo–Anthro (40–48% each)
Total	**10.002**	**50.010**	**86,9%**	

3.2. Preprocessing and Feature Extraction

Each audio sample was processed by PANNs, generating a temporal probability matrix $T \times C$ (T = 501 frames, C = 527 AudioSet classes). Semantic aggregation of the 527 classes into the three ecoacoustic spheres was applied, followed by frame-wise Max Pooling. Max Pooling is preferred over averaging to preserve the dominant acoustic events that drive masking effects, a core dynamic of the Acoustic Niche Hypothesis [Krause 1993; Pijanowski et al. 2011]. Averaging across 527 classes would dilute transient biophonic signals and artificially smooth ecological conflicts in polyphonic environments [Stowell 2022].

The dimensionality reduction from 527 to 3 is intentionally lossy at the micro-event level. Computationally, it acts as a high-variance filter, preventing the Transformer from being overwhelmed by AudioSet-specific fluctuations. Ecologically, it forces the model to focus on the dynamics of competition and masking at the soundscape scale. This approach aligns with the premise that urban environmental health depends on the global interactions between these spheres, rather than on the isolated taxonomy of specific sound sources.

3.3. Transformer Architecture and Training

Temporal modeling was based on an encoder-only Transformer architecture [Vaswani et al. 2017], chosen for its Self-Attention mechanism, which captures temporal dependencies by processing the entire sequence in parallel. The choice of an encoder-only rather than an encoder-decoder framework is justified by the nature of the task: frame-wise classification rather than sequential generation. A decoder would introduce unnecessary parameters and an additional risk of overfitting, given the limited size of the synthetic dataset. The architecture incorporates temporal Positional Encoding to preserve event chronology, followed by two multi-head attention blocks (4 heads, embedding dimension 128).

Training was conducted by minimizing the Sparse Categorical Crossentropy loss function via the Adam optimizer, utilizing Early Stopping (patience = 10) and ReduceLROnPlateau (patience = 5) to mitigate overfitting. Trained exclusively on the Clear Dominance subset (5,001 samples, 25,005 seconds), the process converged at epoch 24. An ablation study (Table 2) confirmed positional encoding's necessity: synthetic gains proved illusory, as real-world deployment (Silent Cities) showed severe performance degradation without temporal context.

Table 2. Positional Encoding Ablation

Configuration	F1-Score (Synthetic)	Spearman Correlation (ρ) with NDSI
With Positional Encoding	0.9960	0.44
Without Positional Encoding	0.9981	0.33

Generalization was evaluated through inference on the 5,001 competition scenes from ESC-50 (unseen during training, but still synthetic). Real-world generalization was subsequently validated across independent datasets. Additionally, the Uniform Manifold Approximation and Projection (UMAP) algorithm [McInnes et al. 2018] was applied to the encoder tensors to visualize the latent space structure, highlighting both the separation between pure classes and the continuity of the learned acoustic transitions.

3.4. Visualization and Geospatial Analysis Methods

Visualization strategies were adapted to each dataset's spatial and temporal characteristics. For purely temporal datasets (Silent Cities [Challéat et al. 2020]), soundscape dynamics were analyzed via time-series evolution of class probabilities, revealing circadian patterns of acoustic dominance shifts. For spatially sparse sensor networks (Urban Soundscapes of the World), Ordinary Kriging [Isaaks & Srivastava 1989; De Coensel et al. 2017] with a spherical variogram was applied to interpolate continuous probability surfaces from point measurements.

Variogram parameters (range, sill, nugget) were automatically estimated via PyKrige. Visual encoding adopted RGB-Alpha mapping Anthropophony (red), Biophony (green), Geophony (blue), with LAeq modulating transparency to simultaneously convey semantic classification and physical sound intensity. For spatio-temporal analysis (SINGA:PURA dataset [Ooi et al. 2021]), successive Kriging was applied across temporal windows (regular grid: 500–1000 m) to generate dynamic 3D surfaces, tracking the expansion and contraction of acoustic dominance zones across

a 24-hour cycle. This approach converts point-based sensor data into continuous, actionable spatial indicators.

4. Results and Discussion

Figure 2 shows that Transformer training converged rapidly (epoch 24), achieving 99.6% accuracy and F1-Score of 0.9960 on the synthetic test set.

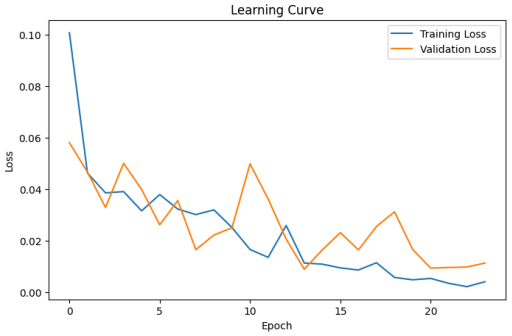


Figure 2. Transformer Learning Curve

In complex acoustic scenarios, metrics based purely on accuracy can be misleading due to dynamic class imbalances. Therefore, performance was measured using the Macro F1-Score, which simultaneously evaluates the harmonic mean of precision and recall across all classes. The framework achieved a Macro F1-Score of 0.9960, demonstrating its strong capacity to distinguish the boundaries between ecoacoustic spheres.

Evaluating the framework against classical baselines (Table 3) substantiates the necessity for deep sequential modeling. Although static models achieved parity in synthetic classification, their lack of chronological sensitivity renders them severely susceptible to predictive fragmentation in real-world continuous streams [Mesaros et al. 2021], causing an abrupt degradation in correlation with the NDSI.

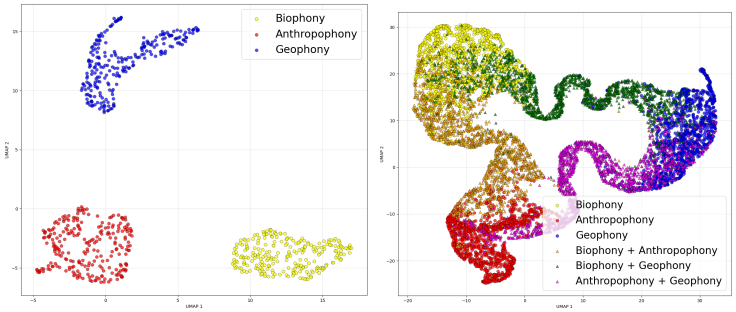
In contrast, sequence-aware models (GRU and Transformer) successfully mitigate this fragmentation. While the Transformer excels in the synthetic task, the GRU achieves slightly higher real-world correlation. This may occur because the GRU's recurrence introduces a smoothing effect over time, producing gradual predictions that align more closely with the continuous NDSI. Conversely, the Transformer's Self-Attention creates sharper decision boundaries, favoring exact event isolation over smooth transitions [Vaswani et al. 2017]. Ultimately, the Transformer is preferred for its superior parallel scalability [Vaswani et al. 2017] and precise masking mitigation prior to Kriging.

Table 3. Performance Comparison with Static Baselines

Model	F1-Score (Synthetic Dataset)	Spearman ϱ (Real Dataset)	Time (s)	Native Temporal Sensitivity
Random Forest	0.9987	0.3519	1.69	No
SVM	0.9987	0.4117	0.07	No
Transformer	0.9960	0.4435	83.02	Yes

GRU	0.9867	0.4838	49.99	Yes

The UMAP projection of the latent space, presented in Figure 3, provides visual evidence of the framework's robustness. Dominant data points organize into three well-defined clusters (Figure 3a), whereas competition scenarios reveal a continuous transitional structure between these groupings (Figure 3b).



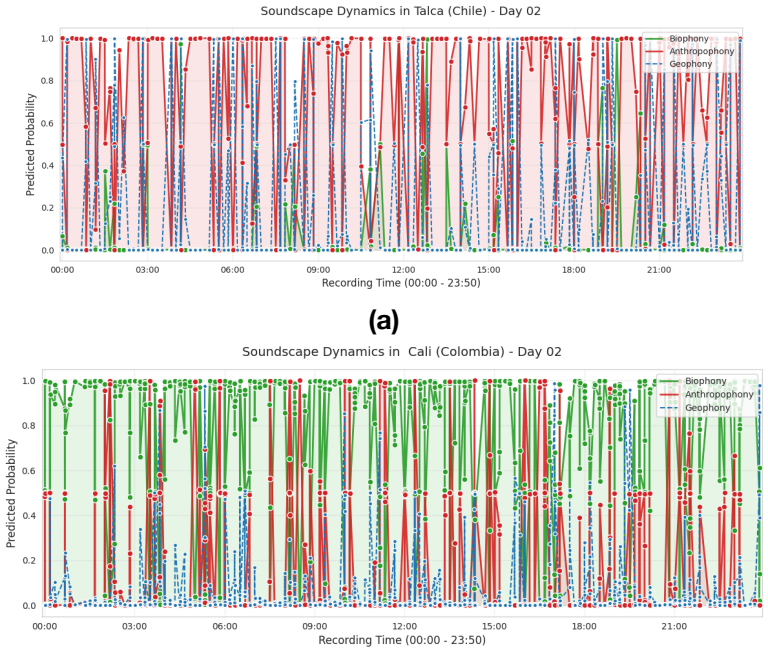
(a) **(b)**

Figure 3. UMAP latent space projection of (a) clear dominance data and (b) conflict/competition data.

This behavior indicates that the model successfully captures the acoustic gradation between sound spheres, rather than merely predicting discrete categories. Consequently, complex urban scenarios can be represented as intermediate combinations of these spheres, enabling the identification of varying levels of biological niche degradation associated with anthropogenic pressure.

4.1. Generalization in Real-World Scenarios

The generalization capacity of the framework was evaluated across three locations from the Silent Cities dataset, representing a land-use gradient: Talca, Chile (100% urban); Cali, Colombia (~94% tree cover); and Jasper, Canada (~34% urban and ~38% tree cover). Figure 4 presents the temporal series for a 24-hour cycle.



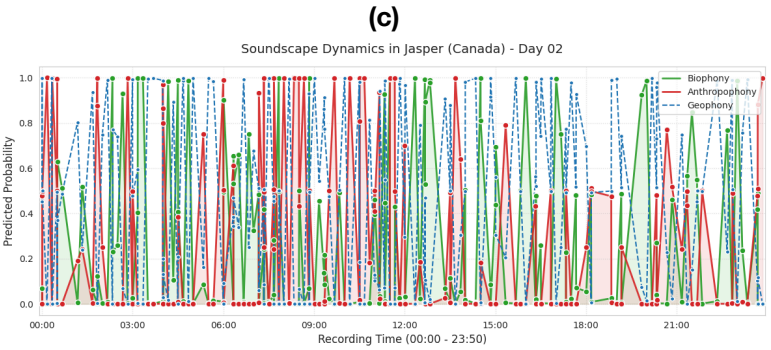


Figure 4. Temporal dynamics of (a) Talca, Chile; (b) Cali, Colombia; and (c) Jasper, Canada.

Generalization across the land-use gradient (Talca: 100% urban; Cali: ~94% tree cover; Jasper: ~34% urban) reveals key patterns (Figure 4). Talca exhibits extreme anthropogenic saturation, while Cali shows biophony dominance interrupted by sparse urban noise. Most compelling is Jasper's temporal niche partitioning, despite balanced land use, biophony reclaims the acoustic niche during dawn chorus and nocturnal periods, suggesting masking effects extend beyond physical urbanization.

Moving beyond temporal analysis to investigate the geographic distribution of sound masking, the application of Ordinary Kriging to the USotW dataset generated continuous probability surfaces, as presented in Figure 5.

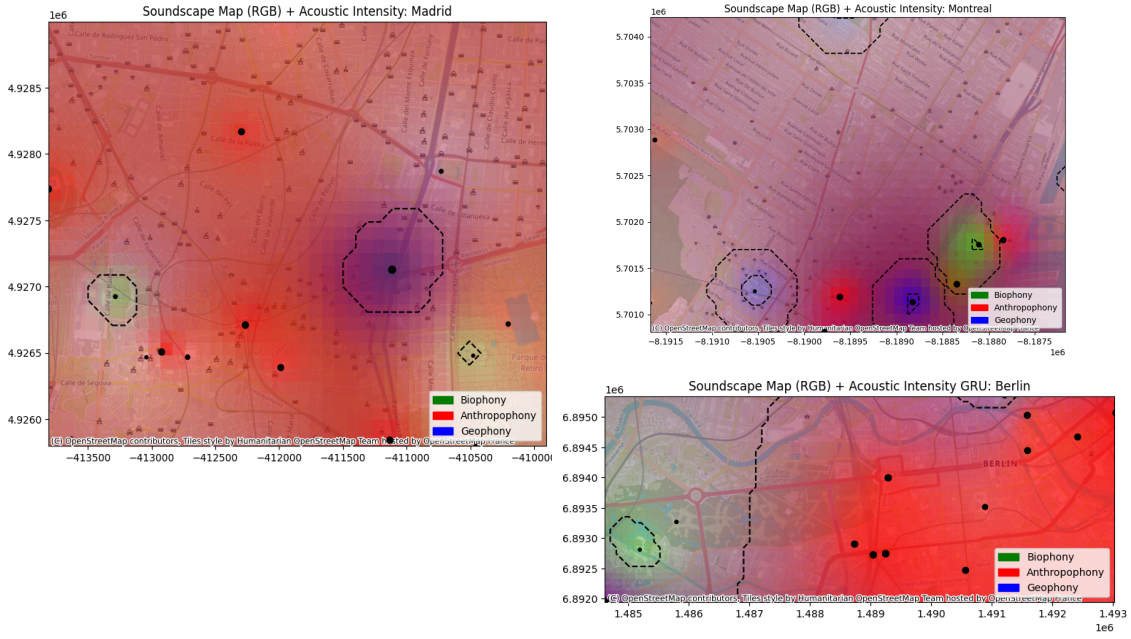


Figure 5. Ordinary Kriging results on the USotW dataset

Three metropolises with distinct morphological profiles were selected: Madrid, Berlin, and Montreal. The RGB-Alpha mapping visualization revealed a strong correlation between the semantic classification and physical sound energy (LA_{eq}). Regions of higher opacity, indicative of greater sound intensity, coincide with zones of Anthropophony (represented in red), particularly in traffic hubs. This reinforces the role

of mechanical sources as the primary vectors of noise pollution. Montreal exhibits acoustic patterns in purplish tones, resulting from the spatial fusion of Anthropophony (red) and Geophony (blue), which indicates the interaction between urban noise and natural events. In Berlin, light green areas map out "islands of Biophony" concentrated within parks, where fauna persists under reduced anthropogenic pressure. The contour lines delineate critical transition zones subjected to intense sound masking, configuring priority boundaries for urban mitigation interventions, such as acoustic barriers and green corridors.

Two validations confirmed real-world generalization: Silent Cities (ρ=0.44, $p<0.001$) and USotW (ρ=0.54, $p<0.001$), as Figure 6 illustrates for the latter. Despite sparse sensors, these correlations are highly relevant in ecoacoustics given extreme acoustic variability and absence of real-world fine-tuning.

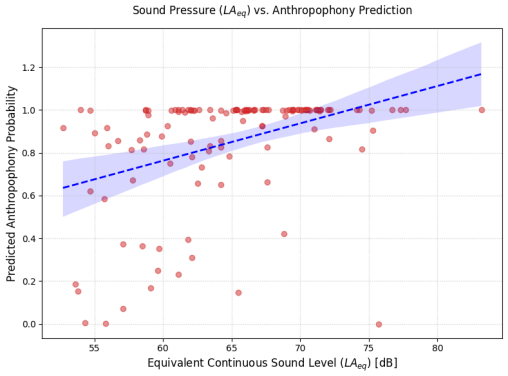


Figure 6. Spearman correlation between LA_{eq} and Anthropophony predictions

Additionally, the quality of the geospatial interpolation was evaluated via spatial Leave-One-Out Cross-Validation (LOOCV) across three metropolises from the USotW dataset, as reported in Table 4. The Root Mean Square Error (RMSE) for Berlin, the highest among the three, is ecologically coherent: the dispersion of Biophony islands within parks creates abrupt spatial gradients. Kriging interpolates these gradients with greater uncertainty given the limited number of deployed sensors.

Table 4. Reported RMSE from spatial LOOCV

City	RMSE	Number of sensors
Madrid	0.0735	16
Montreal	0.2276	11
Berlin	0.2601	14

Finally, the analysis of the SINGA:PURA dataset (Figure 7) incorporates the temporal dimension into the framework. It reveals elevated and persistent Anthropophony in Singapore during the early morning hours (00:00–05:00), reflecting the metabolism of a dense equatorial metropolis. The successive application of Kriging across shifting temporal windows tracked the fluid expansion and retraction of acoustic footprints. This approach provides a dynamic environmental diagnosis that overcomes the spatial and static metric limitations of point-based sensors.

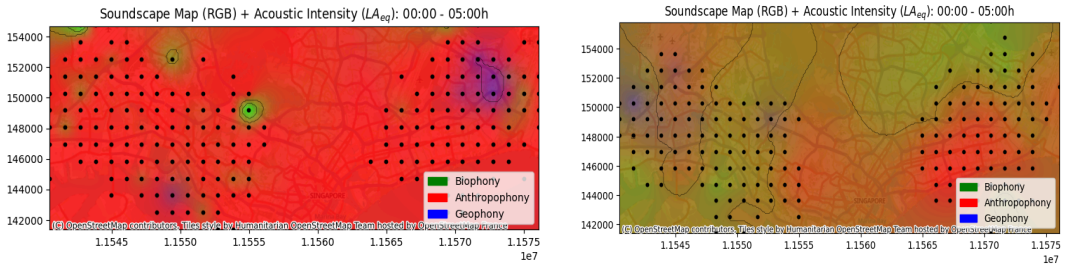


Figure 7. Ordinary Kriging results across time windows on the SINGA:PURA dataset.

4.2. Discussion

The results confirm the central hypothesis of this study: a model trained exclusively on controlled synthetic data can generalize to heterogeneous real-world urban environments, capturing Biophony suppression dynamics with statistically significant correlations across three independent datasets. In comparison, Quinn et al. (2022) employed CNNs trained on real-world data with weak labels, achieving satisfactory performance in ecoacoustic classification but facing limitations when representing transition gradients. In this work, training with precise ground-truth in competition scenes enabled the Transformer to model these transitions more consistently. This result is strongly corroborated by the continuous structure observed in the UMAP-projected latent space.

The case of Jasper warrants special attention: although it presents a relatively balanced land use, its soundscape exhibited a profile similar to that of denser metropolitan areas. This finding suggests that the effects of acoustic masking on Biophony may extend significantly further than physical urbanization indicators imply. Taken together, these findings demonstrate that integrating temporal modeling with geostatistics allows for the conversion of massive volumes of acoustic data into spatial indicators that are highly valuable for environmental monitoring, urban planning, and biodiversity conservation.

5. Conclusion

All three objectives were achieved, with objective (iii) achieved partially: (i) generalization to three real-world datasets ($\rho = 0.54$), (ii) temporal characterization of biophony suppression, including Jasper's niche partitioning, and (iii) spatialization via Kriging, validated by LOOCV (RMSE = 0.07 in Madrid, though higher in cities with sparser sensor networks). Results demonstrate that synthetic training enables robust real-world generalization for urban acoustic monitoring.

The contributions of this study are articulated along three axes: methodological, through a controlled synthetic generation pipeline that produced 10,002 samples with precise ground-truth, made publicly available for reproducibility; ecological, by quantifying the asymmetry predicted by the Acoustic Niche Hypothesis [Krause 1993]; and applied, by converting probabilistic predictions into actionable acoustic dominance surfaces for urban planning. Limitations include the reliance on pseudo-labels via PANNs, the absence of synchronized empirical field validation, and the lack of physical

acoustic propagation modeling. Max Pooling may suppress low-energy events; weighted mean strategies warrant future ablation.

The framework operates at the soundscape level, not fine-grained species classification. Potential edge cases include acoustic masking by structurally similar vocalizations, prosodic ambiguities, or acoustic pareidolia. Rigorous testing against adversarial soundscapes is recommended for sensitivity-critical applications.

Future work will focus on porting the framework to embedded systems coupled with physical sensors for edge computing; adopting dynamic decision margins adaptable to local contexts; and applying the framework to Brazilian biomes, where Passive Acoustic Monitoring (PAM) faces unprecedented challenges in both scale and sound diversity.

References

Challéat, S., et al. (2020). Silent Cities: An acoustic monitoring network of the Covid-19 lockdown. CNRS. https://osf.io/h285u/

Correia, G., & Ventura, T. (2025). Assessing the Potential of Deep Learning in the Acoustic Identification of Threatened Birds in the Pantanal. In Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional, (pp. 1491-1502). Porto Alegre: SBC. doi:10.5753/eniac.2025.12424

De Coensel, B., Sun, K., & Botteldooren, D. (2017). Urban Soundscapes of the World: Selection and reproduction of urban acoustic environments with soundscape in mind. In INTER-NOISE and NOISE-CON Congress and Conference Proceedings (Vol. 255, No. 2, pp. 5407-5413). Institute of Noise Control Engineering.

Isaaks, E. H., & Srivastava, R. M. (1989). An introduction to applied geostatistics. Oxford University Press.

Kong, Q., et al. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2880–2894.

Krause, B. L. (1993). The niche hypothesis. The Soundscape Newsletter, 6, 6–10.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861

Mesaros, A., Heittola, T., Virtanen, T., & Plumbley, M. D. (2021). Sound event detection: A tutorial. IEEE Signal Processing Magazine, 38(5), 67–83. https://doi.org/10.1109/MSP.2021.3090678

Nery, S., Batista, F., & Bezerra, S. (2025). ColabSomGeo: Ferramenta de monitoramento de poluição sonora subsidiada por modelo de aprendizagem. In Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web, (pp. 99-102). Porto Alegre: SBC. doi:10.5753/webmedia_estendido.2025.16410

Nikolenko, S. I. (2021). Synthetic data for deep learning (Vol. 174). Springer International Publishing.

Ooi, K., et al. (2021). SINGA:PURA: A polyphonic urban sound dataset with spatiotemporal context. In 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 463–470). IEEE.

Peralta, A., Santos, E., Xie, J., & Colonna, J. (2024). Recuperação semântica de paisagens sonoras usando banco de dados vetoriais. In Anais do XV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais, (pp. 51-60). Porto Alegre: SBC. doi:10.5753/wcama.2024.2316

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In 2015 IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1–6). IEEE.

Pijanowski, B. C., Villanueva-Rivera, L. J., Slabbekoorn, H., Farina, A., Krause, B. L., Napoletano, B. M., Gage, S. H., & Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. BioScience, 61(3), 203–216. https://doi.org/10.1525/bio.2011.61.3.6

Quinn, C. A., et al. (2022). Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data. Ecological Indicators, 138, 108831.

Ronchini, F., & Serizel, R. (2022). A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes. In ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1031–1035). IEEE.

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279–283.

Stowell, D. (2022). Computational bioacoustics with deep learning: A review and roadmap. PeerJ, 10, e13152.

Sugai, L. S. M., Silva, T. S. F., Ribeiro Jr., J. W., & Llusia, D. (2019). Terrestrial passive acoustic monitoring: Review and perspectives. BioScience, 69(1), 15–25. https://doi.org/10.1093/biosci/biy147

Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).