Subdomain Identification Strategies for Efficient Machine Learning Models*

Samuel R. Torres[1]**,Rocío Zorrilla**[1]**, Raphael Saldanha**[3]**,**
Victor Ribeiro[1]**, Eduardo H. M. Pena**[2]**, Fabio Porto**[1,3]

[1]National Laboratory of Scientific Computing (LNCC)
Caixa Postal 25651-075 – Petrópolis - RJ - Brazil

[2]Federal University of Technology - Paraná,
Campos Mourão - PR - Brazil

[3]INRIA, France

{samuelrt, victorr}@posgrad.lncc.br, {romizc, fporto}@lncc.br

eduardopena@utfpr.edu.br, raphael.de-freitas-saldanha@inria.fr

Abstract. *The performance of machine learning models depends on both the quality of the data and the selection of the models. This work highlights the importance of identifying regions in the input space with similar behavior to improve predictive accuracy. Clustering techniques can partition multivariate time series (MTS) into meaningful subsets, enabling the training of specialized models. We employ k-Medoids and quadtree-based clustering, both using dynamic time warping (DTW) as a similarity measure, and the quadtree also incorporates entropy for partitioning. Long Short-Term Memory (LSTM) networks are trained on these clusters and compared to a global model trained on the entire dataset. The results support the subset modeling hypothesis, showing that models trained in clusters can achieve comparable performance to a global model. This approach offers a comparable alternative that balances prediction accuracy with computational and interpretable advantages.*

1. Introduction

In this work, we extend a proposition from [Montero-Manso and Hyndman 2021], which states that MTS forecasting can be approached in two ways: locally or globally. The local approach involves modeling each MTS, without any grouping, whereas the global approach consists of training a single model on the entire MTS dataset. We leverage the notion of similarity in MTS data that has been addressed using various distance measures, among which DTW, introduced by [Sakoe and Chiba 2003]. Identifying groups of similar MTS enables the adoption of an alternative modeling strategy, such as cluster modeling, that lies between local and global approaches. In that sense, our extension goes beyond the original proposed framework, allowing the development of specialized models for subsets of data that share similar characteristics.

A key consideration is whether a global model, if sufficiently simple and interpretable, is adequate to capture all relevant aspects of the groups of MTS. However, if the

*The authors acknowledge the Brazilian funding agencies CNPq, CAPES and Petrobras Termo de Cooperação 0050.0122040.22.9 for their financial support in the development of this work.

model becomes excessively complex, reducing this complexity by adopting multiple specialized models may be a preferable alternative. To explore this, we hypothesized that the subset modeling approach of [Ribeiro et al. 2023], which identifies subdomains and trains local models, could lead to improved predictions. Our work focuses on meteorological multivariate time series, leveraging the inherent challenges in similarity measurement.

Since similarities between MTS can be quantified, clustering algorithms can be applied to group MTS based on their pairwise similarity. Common approaches include k-Means [MacQueen et al. 1967] and k-Medoids [Kaufman and Rousseeuw 2009], although their reliance on geometric assumptions may introduce bias. Furthermore, we explore a quadtree-based clustering method that uses normalized entropy, inspired by [Angelo 2016] and further extended in [Basu and Sengupta 2015]. For the modeling phase, LSTM networks are used due to their proven ability to capture temporal dependencies in sequential data [Hochreiter 1997].

The results obtained suggest that the proposed strategy achieves error levels comparable to those of the local modeling approach. This finding indicates that it is not necessary to build a model for each MTS. Instead, identifying groups of similar time series emerges as a viable alternative, allowing the development of specialized models that generalize well within each group while significantly reducing the modeling complexity.

The remainder of this paper is structured as follows. Section 2 reviews related works on time series clustering and subset modeling. Section3 introduces key definitions. Section 4 details the proposed methodology, including the clustering strategies and the modeling framework. Section 5 presents the experimental setup and the results. Finally, Section 6 discusses the main findings and outlines future research directions.

2. Related Work

One of the works that inspired our use of similarity-based clustering for MTS is [Singhal and Seborg 2005], in which the authors proposed a methodology that leverages the degree of similarity as a key parameter to cluster multivariate time series.

The conceptual framework for distinguishing between local and global approaches was established by [Montero-Manso and Hyndman 2021]. The authors cast the problem of forecasting a set of MTS between two extremes, determined by how the set is partitioned. Partitioning the set into individual MTS leads to the local approach, whereas the global approach operates on the trivial partition that maintains the set as a whole. Furthermore, they demonstrate that the complexity of a learning algorithm can be controlled by the granularity of the partitioning. This result emphasizes the inherent trade-off in adopting more sophisticated partitioning strategies, such as data-driven partitioning, which encompasses clustering and its fuzzy weighted alternatives.

Another work proposed a framework to reduce computational costs while maintaining adequate model accuracy[Ribeiro et al. 2023]. The success of machine learning (ML) systems is largely dependent on the availability, volume, quality, and efficient computational resources of the data. To address these challenges, the authors suggest identifying "subdomains" within the input space and training local models that produce better predictions for samples from these specific subdomains, rather than relying on a single global model trained on the full dataset. Their experimental evaluation, conducted on

two real-world datasets, demonstrates that subset modeling (i) improves predictive performance compared to a global model, and (ii) enables more data-efficient training.

In [Zorrilla Coz 2021], a hybrid approach is proposed that combines per-element and cluster-based modeling for univariate time series. By generating models only for representative elements and using a time series classifier to generalize across the domain, it achieves predictive accuracy comparable to conventional methods with significantly lower computational cost. Clustering is used as a structural guide, while classification ensures adaptability to local variations that pure clustering may miss.

The difference from these related works is the focus on multivariate time series, rather than tabular data or univariate time series. In the following section, we define the key concepts that support this work.

3. Background

This work focuses on analyzing a set of multivariate time series to identify regions with similar temporal patterns. The underlying motivation is to capture these regions as coherent groups that exhibit similar temporal behavior. To capture these similarities, we used DTW as the primary similarity metric. A clear understanding of our approach requires the introduction of several foundational concepts. The authors [Cao and Liu 2016] define an MTS as follows:

Definition 3.1 *Consider multivariate time series X where $X = (x_1^m, x_2^m, \ldots, x_n^m), \quad i = 1, 2, \ldots, n$, and $m \in \mathbb{N}, x_i^m \in \mathbb{R}^m$. Specifically, if $m = 1$, the given series is a univariate time series, and if $m > 1$, it is an MTS.*

The partitioning problem is addressed using two distinct methods: k-Medoids and quadtree. The k-Medoids algorithm is used in combination with DTW as a similarity measure, allowing effective clustering of time series by accounting for temporal distortions and misalignments in the sequences.

For the quadtree clustering method, we incorporate the concept of *normalized entropy* as a criterion for subdivision. This measure allows evaluating the degree of variability or disorder in a group of MTS. The goal is to refine partitions only when the internal heterogeneity is high, preventing unnecessary subdivisions.

Definition 3.2 *The normalized entropy $\mathcal{H}$ is defined as*

$$\mathcal{H} = -\frac{\sum_{i=1}^{N} \omega_i \log_2(\omega_i)}{\log_2(N)} = \frac{\mathbb{S}_{\mathbb{P}}}{\mathbb{S}_{max}},$$

where ω_i denotes the DTW distance between two multivariate time series (MTS), $\mathbb{S}_{\mathbb{P}}$ is the Shannon entropy of the distribution, N be the set of off-diagonal upper triangular elements of the distance matrix, and $\mathbb{S}_{max} = \log_2(N)$ represents the maximum possible entropy.

In our methodology, the quadtree algorithm halts further space division when the normalized entropy of a node falls below a predefined threshold and a minimum number of samples is retained. This ensures the formation of meaningful and compact subdomains with similar temporal behavior.

4. Methodology

The methodology developed in this work extends the perspective by [Montero-Manso and Hyndman 2021], who address forecasting problems. They propose two main strategies: the **local approach**, where each MTS is modeled under the assumption that it may originate from a distinct data-generating process; and the **global approach**, which assumes a shared structure and trains a single model in all MTS. We introduce an intermediate strategy, the **cluster approach**, where MTS are first grouped using clustering techniques, and then separate models are trained for each group. This approach seeks to balance the trade-off between model complexity and generalization by exploiting similarities within clusters. Our methodology assumes that some MTS are similar enough to be modeled jointly. This leads to a **cluster approach** modeling strategy, where subsets of series with similar temporal behavior are identified and modeled. This balances the generalization ability of global models with the specialization of local models. By leveraging intra-cluster similarities, the approach aims to improve forecasting accuracy and computational efficiency without the need to build a separate model for each MTS.

4.1. Partitioning process

In the case of k-Medoids, the calculation requires a measure of similarity between pairs of elements in the data set [Park and Jun 2009]. This similarity is quantified using DTW, which is particularly well suited for comparing MTS data. To evaluate the quality of the resulting clustering, the silhouette score was used. The second partitioning method used in this work is the quadtree algorithm. Here, the standard quadtree[Finkel and Bentley 1974] is adapted to recursively divide the input space D into four disjoint regions, producing a hierarchical partition. Unlike a standard quadtree that relies solely on a minimum number of elements as the stopping criterion, the proposed method also considers the normalized entropy of pairwise similarities. Recursion stops when either condition is satisfied. The quality of the quadtree partition is assessed on the basis of its effectiveness in reducing entropy.

4.2. Modeling Strategies

Our modeling strategy consists of defining a fixed temporal division that will be used for model training. Specifically, this division is the 80%/20% (training/test) of the total length of each MTS. It is important to note that all MTSs have the same temporal length.

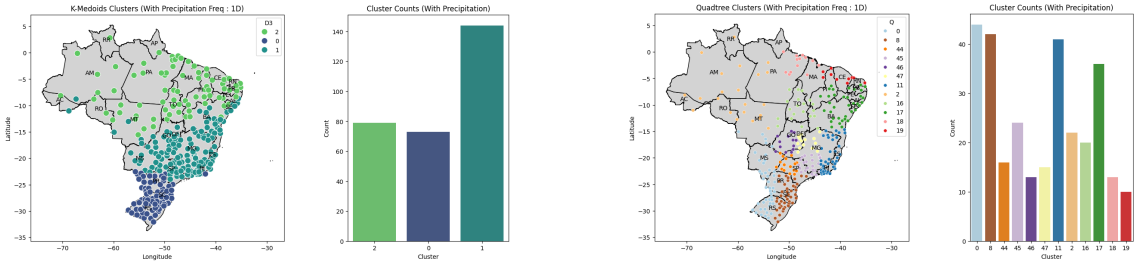
In this part, we consider three modeling approaches. In the local modeling strategy, one model is trained for each MTS, resulting in the same number of models as there are MTS. For the cluster modeling approach, a model is trained for each cluster, leading to a number of models equal to the number of identified clusters. Finally, in the global modeling approach, a single model is trained using the entire set of MTS.

5. Experiments and Results

The data set used was obtained from [Instituto Nacional de Meteorologia 2024], which provides organized yearly meteorological data from various automatic weather stations (296). For this work, we used data from 2017 and 2018. The variables analyzed include pressure (P, in hPa), temperature (T, in C), relative humidity (H, in %), wind speed (W,

in m/s), and precipitation (Prec, in mm). Although latitude and longitude were also recorded as part of the MTS, they are not used in the k-Medoids analysis and serve only to construct the quadtree grid. The model aims to predict the accumulated precipitation for the next day based on data from the previous seven days.

The Figure 1 shows the clusters obtained, three clusters by k-Medoids, and twelve for quadtree. The k-Medoids method was executed varying $k = 3, \ldots, 10$, and we select $k = 3$ because it generates a better silhouette score, in addition the execution time takes approximately 354.019 seconds. In contrast, the quadtree method with the experimental hyper-parameters $\mathcal{H} = 0.5$ and $\min\_points = 60$, completed the clustering process in 140.543 seconds. This shows that the quadtree approach is significantly more efficient in terms of computational time.



(a) Clustering by k-Medoids **(b) Clustering by quadtree**

Figure 1. Comparison of the obtained clusters using precipitation.

To properly interpret the following Table 1, it is necessary to define the key terms used. The **Global Error** refers to the error incurred by the global model when evaluated in the test data, measured by the Root Mean Squared Error (RMSE), which in this case is **11.093** (no clustering). The **Global Region Error** denotes the error produced by the global model when evaluated on the test data corresponding to the cluster region. The **Cluster Error** represents the error incurred by the cluster-specific model when evaluated on its respective test data. Finally, the **Local Error** refers to the error produced by the local model when tested on its own data test.

Method	Cluster ID	Global Region Error	Cluster Error	Local Error
k-medoids	0	13.230	13.708	12.022
	1	10.437	10.275	10.170
	2	10.822	10.236	8.835
quadtree	0	13.104	12.209	11.778
	2	15.587	12.150	13.533
	8	14.391	12.311	11.283
	11	10.417	9.342	9.503
	16	12.275	11.089	10.953
	17	6.500	5.567	5.136
	18	12.862	9.603	9.008
	19	8.263	5.597	4.898
	44	12.738	11.792	11.011
	45	11.924	11.045	11.387
	46	11.858	11.025	11.334
	47	13.658	12.587	12.916

Table 1. Comparison of RMSE errors for the global model and local models trained on clusters defined by k-medoids and quadtree. The *Local Error* corresponds to the average prediction error within each cluster.

Table 2 presents the average relative reductions in execution time required to train each model, taking the global model as the baseline. The k-Medoids-based approach achieved a time reduction of approximately **48.99%**, while the quadtree method produced a more substantial reduction of approximately **85. 24%**. Finally, training the models for each MTS resulted in the greatest efficiency gain, with an approximate reduction of **99.04%** compared to the global model. These results suggest that, under a parallelized execution framework, such reductions would directly translate into equivalent gains in computational efficiency, as the localized models could be trained simultaneously.

Time Step	Global Model (s)	k-Medoids (s)	quadtree (s)	Local (s)
1 Day	762.46	389.06	112.51	7.34

Table 2. Execution time for fitting the models

6. Conclusions and Future Works

The results show that the error achieved by the cluster approach is comparable to that of the local strategy. This indicates that it is not necessary to train a separate model for each MTS. Instead, grouping time series based on similar temporal behavior yields competitive results. Specifically, in our case study that involved automatic weather stations in Brazil, the findings suggest that it is unnecessary to build a model for each station. The grouping of stations according to their temporal patterns offers an effective alternative, achieving similar predictive performance while significantly reducing the number of required models.

Latitude and longitude were excluded from the cluster analysis using k-Medoids, as their proximity could unduly influence cluster formation. The primary objective in this case was to capture temporal patterns rather than spatial relationships. In contrast, quadtree experiments required the use of latitude and longitude to define the spatial grid in which clusters are identified. In this context, geographic coordinates are not analyzed as time-dependent variables but serve as a necessary spatial reference for implementing the quadtree algorithm.

These findings reinforce the notion that leveraging spatial or temporal similarity through appropriate clustering, especially when guided by spatial structures such as quadtrees, can lead to more specialized models that better capture local dynamics. Our work provides evidence that structured clustering not only improves prediction accuracy but also offers a more efficient and interpretable modeling strategy.

Future research should evaluate alternative criteria for quadtree subdivision to improve clustering quality and explore adaptive strategies. Furthermore, systematic hyperparameter optimization could improve the model performance between clusters. Finally, optimizing the quadtree algorithm for scalability is essential to handle larger datasets efficiently.

References

Angelo, A. (2016). A brief introduction to quadtrees and their applications. In *Style file from the 28th Canadian Conference on Computational Geometry*.

Basu, D. and Sengupta, S. (2015). A novel quad tree based data clustering technique. In *2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pages 157–162. IEEE.

Cao, D. and Liu, J. (2016). Research on dynamic time warping multivariate time series similarity matching based on shape feature and inclination angle. *Journal of Cloud Computing*, 5(1):11.

Finkel, R. and Bentley, J. (1974). Quad trees: A data structure for retrieval on composite keys. *Acta Inf.*, 4:1–9.

Hochreiter, S. (1997). Long short-term memory. *Neural Computation MIT-Press*.

Instituto Nacional de Meteorologia (2024). Dados históricos - inmet. Accessed: 2024-06.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Montero-Manso, P. and Hyndman, R. J. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, 37(4):1632–1653.

Park, H.-S. and Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341.

Ribeiro, V., Pena, E. H., Saldanha, R., Akbarinia, R., Valduriez, P., Khan, F. A., Stoyanovich, J., and Porto, F. (2023). Subset modelling: A domain partitioning strategy for data-efficient machine-learning. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 318–323. SBC.

Sakoe, H. and Chiba, S. (2003). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Singhal, A. and Seborg, D. E. (2005). Clustering multivariate time-series data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(8):427–438.

Zorrilla Coz, R. M. (2021). *A Spatial-Temporal Aware Model Selection for Time Series Analysis*. PhD thesis, Laboratório Nacional de Computação Científica, Petrópolis, RJ, Brasil. Thesis for the degree of Doctor of Sciences in Computational Modeling.