Subset Models for Multivariate Time Series Forecast

Raphael de Freitas Saldanha, Victor Ribeiro, Eduardo Peña, Marcel Pedroso, Reza Akbarinia, Patrick Valduriez, Fabio Porto

Subset Models for Multivariate Time Series Forecast

Raphael Saldanha*, Victor Ribeiro† Eduardo H. M. Pena‡, Marcel Pedroso§,
Reza Akbarinia*, Patrick Valduriez*†, and Fabio Porto†
*Inria, University of Montpellier, CNRS, LIRMM, Montpellier, France
†DEXL, LNCC, Petrópolis, Brazil
‡DACOM, UTFPR, Campo Mourão, Brazil
§PCDaS, LIS, ICICT, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

***Abstract*—Multivariate time series find extensive applications in conjunction with machine learning methodologies for scenario forecasting across various domains. Nevertheless, certain domains exhibit inherent complexities and diversities, which detrimentally impact the predictive efficacy of global models. This ongoing study introduces a Subset Modeling Framework designed to acknowledge the inherent diversity within a domain's multivariate space. Comparative assessments between subset models and global models are conducted in terms of performance, revealing compelling findings and suggesting the potential for further exploration and refinement of this novel framework.**

***Index Terms*—domain diversity, subsets, machine learning, dengue, climate.**

I. Introduction

Time series forecasts may take advantage of multivariate abundant data that are available about certain phenomena by adopting machine learning techniques suitable for multivariate data. However, due to the domain's intrinsic characteristics of complexity and diversity, Machine Learning Systems (MLS) may not perform equally well forecasting different parts of the input, despite presenting a good overall performance. Our ongoing work endeavors to address this challenge by exploring "subsets" within the data domain in a multivariate space for model training and forecasting.

As a motivating example, we consider the task of forecasting dengue disease incidence in Brazilian cities using climate indicators as covariates. Dengue disease has been endemic in Brazil since 1986, presenting 2 million cases only in 2022 and a clear tendency to increase in the next years, causing a heavy health burden over the population [1]. The disease is transmitted to humans by the bite of infected mosquitos from the *Aedes* genus carrying the dengue virus (DENV) variants. Thus, the disease spread is affected by the circulation of the virus in the human population and the infestation of competent disease vectors. The mosquito's life cycle is deeply affected by climate factors, such as precipitation (to enable breeding sites) and temperature (with thresholds for its reproduction, feeding, and biological activity). Hence, those indicators are usually admitted as covariates in a statistical modeling approach [2]–[4].

We introduce a modeling framework that accommodates shared features characteristics and regional variations across diverse units (e.g. municipalities), offering cost-effective training and robust prediction capabilities. The concept is as follows: (1) identify subsets within the dataset with similar covariate patterns and train models specific to each subset; (2) map incoming samples to the appropriate subset, as a function of the similarity between the data distributions; and (3) use the model trained on the subset data for prediction.

A. Summary of contributions

This work introduces: (i) a subset machine learning model construction approach based on multivariate time series data; and (ii) an experimental evaluation of the proposed approach that shows the viability of its application on a real case scenario of multivariate modeling a climate-sensitive disease. The rest of this paper is organized as follows. In Section II, we detail our framework. In Section III, we empirically demonstrate its effectiveness by forecasting dengue case incidence in Brazil, and in Section (IV) we compare our approach with related work.

II. Subset Modeling Framework

The objective of this modeling framework is to improve forecasting accuracy by training different models based on subsets of the data domain, in comparison with the accuracy obtained by a global model that was trained on all data. Figure 1 depicts a pipeline for the subset-based model life-cycle implementation.

Departing from the traditional model life-cycle, this pipeline introduces the identification of domain data subsets and their use in model training and forecasting. A data domain may be split into multiple partitions of different sizes and composition. The ideal subsets' configuration is the one that maximizes its model accuracy in contrast with a global model.

Those subsets may be defined based on previous knowledge about the domain or by a data-driven approach, which is the focus of this paper. Specifically, our definition of a subset is a group of data points (i.e. multivariate time series subsequences) from a data domain that present similar characteristics that can be measured and enhance the model's forecasting performance.

A. Definition

We consider a dataset D whose domain is a set of time series (ts), ordered by time, and having some measurement values at each time instant.

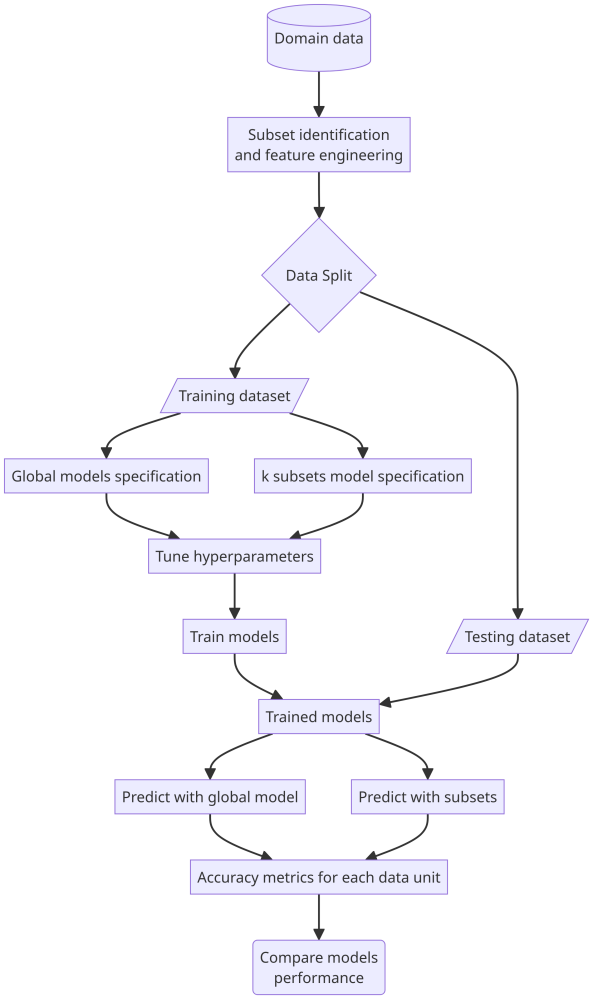

Fig. 1. Subset Modeling pipeline

Let $P_i = \{S_1, S_2, \ldots, S_k\}$ be a partitioning of D, such that $S_j = \{s_{j1}, s_{j2}, \ldots, s_{jn}\}$ is a subset of time series from D, and $S_j \cap S_t = \emptyset$ if $j \neq t$, and $D = \cup_{j=1..k}\, S_j$. For each subset $S_j \in P_i$, we train a machine learning model, using a given *learner* algorithm. Each machine-learning model m_j in M, $1 \leq j \leq k$, is built on the samples of the corresponding subset $S_j \in P_i$. We call P_i a partitioning of dataset D. We want to find a partitioning P_i of the dataset D, such that when used for training produces models in M with minimum prediction error $E(s)$:

$$P_i := argmin_{P_i \in P} \sum_{j=1..k} E_s((ts_l \in S_{ij}) : m_{ij}(ts_l)) \quad (1)$$

Solving Equation 1 is not practical for large datasets as the number of possible partitions is exponential to the number of time series.

Thus, we propose a data-driven approach based on the k-medoids clustering algorithm that approximates the optimal partitioning in Equation 1 to one exhibiting clusters sharing similar time series.

B. Subsets identification by clustering

Our approach to computing domain data subsets leverages the idea of computing clusters of time series using time series distance metrics. Dynamic Time Warping (DTW) is a distance measure between time series to compare the similarity between their shapes, introduced by [5]. Unlike the Euclidean distance, DTW allows one-to-many points alignment, with the advantages of capturing similarities between pairs of time series of equal or different sizes, and supporting stretching and bending on the time axis [6]. Despite its higher computational cost compared to the Euclidean distance, its flexibility has led to widespread adoption. It is employed in our framework to identify data subsets [7].

The DTW distance was initially proposed to operate on a single-dimensional time series. The generalization of DTW to a multidimensional case is possible by two particular approaches. The *Independent* approach is based on the cumulative distance of all dimensions that are independently measured under DTW, while the *Dependent* approach computes the multidimensional DTW forcing that all dimensions warp identically in a single warping matrix. Complete definitions of those approaches are available at [8]. In this work, the Independent approach is adopted due to its simplicity. After the subset identification, further feature engineering steps are applied to create time-lagged variables from the interest variable and covariates.

C. Modeling

The modeling phase comprehends a series of model specifications to be trained and evaluated. Two global models are built using the full training dataset: a global model considering the dataset with all domain features and a second global model, that contains all domain features and also a feature with the subset-id of the times, in one-hot encoding. Finally, for each time series subset, a model is trained using the data comprised of all time series in that subset.

All models (global and subsets) hyperparameters are tuned, and the models' accuracies are obtained with the testing dataset for comparison and evaluation.

III. Experimental Evaluation

In this section, we evaluate the performance of the proposed Subset Modeling Framework for forecasting cases in the dengue disease incidence domain.

A fairly common prediction model specification for dengue disease cases incidence forecast includes, as predictors, time-lagged values of dengue cases and climate indicators [9]–[11].

In this case, a global model approach is expected to face difficulties, as various locations exhibit distinct temporal and spatial disease transmission patterns, which are influenced by different climate regimes as well as diverse population and urban conditions [12]. Thus, a single-model approach will likely not capture this heterogeneity, as global models approximate the data distribution seen during training and suffer when such distribution exhibits important variations along the domain. Also, building separate (local) models for each

municipality presents challenges in dealing with the scarcity of cases that will not generalize well when deployed, missing the opportunity to learn variations on the data distribution, enabling model generalization from seen data. The proposed subset approach is between the two extremes.

By computing subsets of the domain, we build models that better approximate the dengue case incidence and its predictors.

A. Datasets

Dengue cases dataset. This dataset presents the count of confirmed dengue virus (DENV) cases per epidemiological week and municipality in Brazil. DENV-suspected cases are tracked nationwide by a health information system ("Sistema de Informação de Agravos de Notificação") maintained by the Brazilian Health Ministry. For this research, we computed the number of confirmed dengue cases on each epidemiological week by municipality of residence, from 2011 to 2020. Due to the count data sparsity, we only considered cities with more than 100,000 inhabitants in 2020. Thus, the dataset presents 495 measurements of dengue cases in time, for 333 cities in total.

Climate indicators dataset. This dataset presents average maximum temperature (C degrees), average minimum temperature (C degrees), and total precipitation (mm) indicators per epidemiological week and municipality in Brazil. The indicators were computed using data from the ERA5-Land Reanalysis [13] by computing spatial zonal statistics considering the municipalities boundaries (averages for temperature and sum for precipitation) [14].

B. Experimental setup

The dengue cases and the climate datasets were merged by the municipality's unique identification code, year, and epidemiological week references. Considering the diversity of units and magnitude values among the time series of each indicator, the time series were standardized (*z-score*) with zero mean and one standard deviation to not bias the similarity metric, as proposed by [15].

To determine the subsets, the DTW distance was computed in the multivariate form (considering the dengue cases and climate indicators) and clustered from $k = 3$ to $k = 20$ clusters, being selected the k value presenting the higher silhouette statistic [16], [17].

A sliding window of length $w = 6$ and stride 1 on the complete training sequence was computed for each indicator [18], producing $n - (w + 1)$ subsequences as input for the training process. For each week w, the prediction infers the next measurement value.

The dataset was split into training and testing, using the first (temporally) eight years for training (80%) and testing on the remaining two years (20%).

The random forest learner was adopted due to its capacities on time series forecasting [19], [20], with hyperparameters tuning [21], [22]. The global models and subsets models were trained with the training dataset, and the model's performance was evaluated with the testing dataset. We use the Root Mean Squared Error (RMSE) as the evaluation metric.

C. Results

The DTW multivariate clustering was applied to identify subsets in the dataset domain. Several subsets by varying k from 3 to 20 were tested. The case for $k = 5$ presented the highest silhouette statistic (0.1280), with the following partition (g) sizes: $g_1 = 69$, $g_2 = 62$, $g_3 = 82$, $g_4 = 102$ and $g_5 = 18$. Those partitions include municipalities that are not necessarily neighbors spatially, as illustrated by Figure 2.

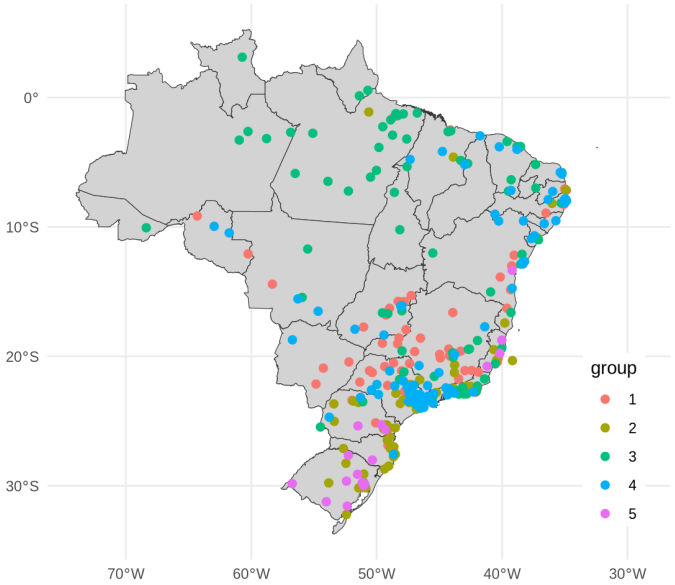

Fig. 2. DTW partition units map

As shown in Figure 2, the partitions present some spatial relationships linked to climate regimes and dengue spread similarities, as expected. For example, the partition g_4 includes municipalities from Southeast, Midwest, and Northeast regions. Those regions have distinct climate regimes in-between, but the municipalities present some similarities in the dengue cases time series.

Considering the five partitions as data subsets for model training and assessment, Figure 3 presents models' accuracy results on testing for the subsets models and global models.

As observed in Figure 3, the global models with and without the partition identification variable (models `global` and `globalID`, respectively) do not present distinguishable RMSE values, implying that the presence of the identification variable at model specification does not affect the global model results, at least for the Random Forest learner.

On the other hand, the subset models present some interesting overall results, especially when considering the size of the training data for each subset model. The models trained on the smaller partitions (g_1, g_2, and g_5) present higher RMSE medians than the global models, implying that the subset models' errors for the municipalities present on those partitions may be similar or higher than the global models.

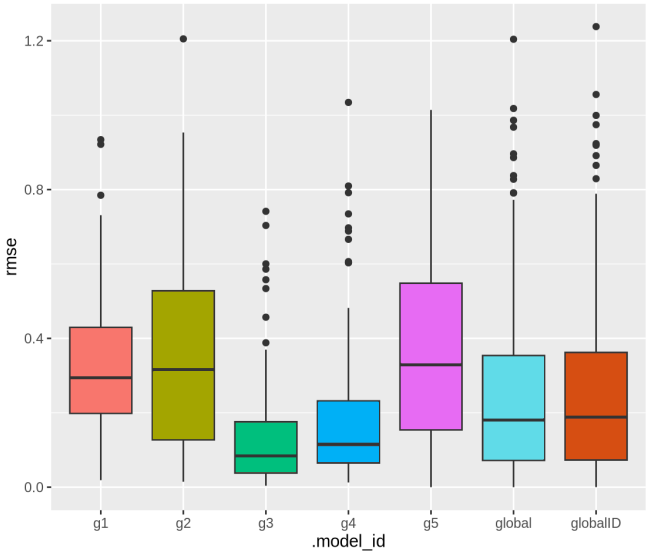

Fig. 3. Performance of subset models vs. global models

On the other hand, the RMSE's medians observed on the bigger partitions (g_3 and g_4) are lower in comparison with the global model, which implies that the proposed approach presents overall better results in comparison with the global model approach for the municipalities on those partitions.

This performance difference between the subsets models is likely related to the number of data points present on each subset, as the larger partitions (g_3 with 82 data points, and g_4 with 102 data points) presented an average better performance in comparison with the others. Anyway, the partitions g_3 and g_4 contain only a fraction of the data used on the global models, indicating that the model performance is affected not only by the partition size but also by its composition.

More specifically, by observing the error for each municipality on its subset model and at the global model, the subset models presented smaller RMSE results at 116 municipalities, implying a model accuracy improvement in 34.83% of the municipalities. For example, the municipality of *São Leopoldo, RS*, presented a RMSE value of 0.48 at the global model (`global`) and 0.28 at the subset model (g_5).

These results from one experiment point that the proposed Subset Modeling Framework can contribute to enhancing the prediction performance for a fraction of the data points, in comparison with the global model approach.

IV. Related work

The work by [23] proposes using time series clustering to create subsets to improve dengue forecast performance. The authors adopt a *k-means* clustering procedure based on time series features, as proposed by [24], being applied to univariate time series of dengue cases. More recently, [25] investigates a similar approach to create subsets to improve model accuracy, with experimental evaluation on dengue and employment datasets. As a clustering technique to identify the subsets, the Dynamic Time Warping distance was used. The approach considers univariate time series of dengue cases.

We advance on these works by specifically addressing and formalizing the subsets approach in a multivariate setting, investigating the model accuracy improvement observed in the presence of covariates for clustering and predicting.

V. Conclusion

In this paper, we proposed a subset strategy for multivariate time series to improve model forecasting accuracy. The subsetting strategy of the time series, the number of subsets, and the subset size play an important role in the model accuracy improvement, as the number and sizes of the subsets affect the training process and the model's performance.

Although the subsets' models' overall performance can be characterized in terms of the RMSE median and other summary statistics, it is important to observe and compare the individual RMSE error obtained on each unit (municipality in our case example) on the different model's training. As it was observed, individual units that were trained within a subset that presented a worse overall performance may still present better accuracy performance in comparison with the global model.

Further work may investigate other approaches for creating the subsets of multivariate time series and studying the impact on the model's performance, including (1) constraints on partition sizes and the total number of partitions; (2) strategies for partitioning multivariate time series, including feature-based approaches [26]; (3) partitioning based on socio-demographic indicators that are related to dengue disease spread; and (4) investigating the performance difference of different learners over the subsets and global models.

Our experimental findings suggest that a subset modeling derived from a multivariate clustering approach can improve the model's predictive performance in comparison with global models for at least a fraction of the units.

VI. Acknowledgement

This work was publicly funded through ANR (the French National Research Agency) program with the reference ANR-23-CE20-0020-01 (DeepPep project) and the CNPq Productivity Fellowship process number 312100/2021-3.

References

[1] Ministério da Saúde, "Monitoramento das arboviroses urbanas: semanas epidemiológicas 1 a 35 de 2023," Tech. Rep. 54, Ministério da Saúde, 2023. Boletim epidemiológico.

[2] L. M. Stolerman, P. D. Maia, and J. Nathan Kutz, "Forecasting dengue fever in Brazil: An assessment of climate conditions," *PLoS ONE*, vol. 14, Aug. 2019.

[3] R. Lowe, B. Cazelles, R. Paul, and X. Rodó, "Quantifying the added value of climate information in a spatio-temporal dengue model," *Stochastic Environmental Research and Risk Assessment*, vol. 30, pp. 2067–2078, Dec. 2016.

[4] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, "Forecast of Dengue Incidence Using Temperature and Rainfall," *PLoS Neglected Tropical Diseases*, vol. 6, Nov. 2012.

[5] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," *AAAIWS'94: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 359–370, 1994.

[6] D. Cao and J. Liu, "Research on dynamic time warping multivariate time series similarity matching based on shape feature and inclination angle," *Journal of Cloud Computing*, vol. 5, p. 11, Dec. 2016.

[7] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – A decade review," *Information Systems*, vol. 53, pp. 16–38, Oct. 2015.

[8] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing DTW to the multi-dimensional case requires an adaptive approach," *Data Mining and Knowledge Discovery*, vol. 31, pp. 1–31, Jan. 2017.

[9] X. Y. Leung, R. M. Islam, M. Adhami, D. Ilic, L. McDonald, S. Palawaththa, B. Diug, S. U. Munshi, and M. N. Karim, "A systematic review of dengue outbreak prediction models: Current scenario and future directions," *PLOS Neglected Tropical Diseases*, vol. 17, p. e0010631, Feb. 2023.

[10] M. Cabrera, J. Leake, J. Naranjo-Torres, N. Valero, J. C. Cabrera, and A. J. Rodríguez-Morales, "Dengue Prediction in Latin America Using Machine Learning and the One Health Perspective: A Literature Review," *Tropical Medicine and Infectious Disease*, vol. 7, Oct. 2022.

[11] W. Hoyos, J. Aguilar, and M. Toro, "Dengue models based on machine learning techniques: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 119, Sept. 2021.

[12] C. Barcellos and R. Lowe, "Expansion of the dengue transmission area in Brazil: the role of climate and cities," *Tropical Medicine & International Health*, vol. 19, no. 2, pp. 159–168, 2014.

[13] R. Saldanha, R. Akbarinia, M. Pedroso, V. Ribeiro, C. Cardoso, E. H. M. Pena, P. Valduriez, and F. Porto, "Zonal statistics datasets of climate indicators for Brazilian municipalities," *Environmental Data Science*, vol. 3, p. e2, 2024.

[14] C. Tomlin, "Map algebra: one perspective," *Landscape and Urban Planning*, vol. 30, pp. 3–12, 10 1994.

[15] E. Keogh and S. Kasetty, "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration," *Data Mining and Knowledge Discovery*, vol. 7, pp. 349–371, 2003.

[16] A. Sardá-Espinosa, "Time-series clustering in r using the dtwclust package," *The R Journal*, 2019.

[17] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, pp. 243–256, Jan. 2013.

[18] R. Lowe, S. A. Lee, K. M. O'Reilly, O. J. Brady, L. Bastos, G. Carrasco-Escobar, R. De Castro Catão, F. J. Colón-González, C. Barcellos, M. S. Carvalho, M. Blangiardo, H. Rue, and A. Gasparrini, "Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: A spatiotemporal modelling study," *The Lancet Planetary Health*, vol. 5, pp. e209–e219, Apr. 2021.

[19] M. J. Kane, N. Price, M. Scotch, and P. Rabinowitz, "Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks," *BMC Bioinformatics*, vol. 15, p. 276, Dec. 2014.

[20] R. P. Masini, M. C. Medeiros, and E. F. Mendes, "Machine learning advances for time series forecasting," *Journal of Economic Surveys*, vol. 37, pp. 76–111, Feb. 2023.

[21] M. Kuhn, "Futility Analysis in the Cross-Validation of Machine Learning Models," May 2014.

[22] M. Kuhn, *finetune: Additional Functions for Model Tuning*, 2023. https://github.com/tidymodels/finetune, https://finetune.tidymodels.org.

[23] J. V. Bogado, D. H. Stalder, C. E. Schaerer, and S. Gómez -Guerrero, "Time Series Clustering to Improve Dengue Cases Forecasting with Deep Learning," in *Proceedings - 2021 47th Latin American Computing Conference, CLEI 2021*, Institute of Electrical and Electronics Engineers Inc., 2021.

[24] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *Expert Systems with Applications*, vol. 140, p. 112896, Feb. 2020.

[25] V. Ribeiro, E. H. M. Pena, R. Saldanha, R. Akbarinia, P. Valduriez, F. A. Khan, J. Stoyanovich, and F. Porto, "Subset Modelling: A Domain Partitioning Strategy for Data-efficient Machine-Learning," in *Anais Do XXXVIII Simpósio Brasileiro de Banco de Dados (SBBD 2023)*, (Brasil), pp. 318–323, Sociedade Brasileira de Computação - SBC, Sept. 2023.

[26] M. Tadayon and Y. Iwashita, "A clustering approach to time series forecasting using neural networks: A comparative study on distance-based vs. feature-based clustering methods," Jan. 2020.