2019 IEEE International Conference on Big Data (Big Data)

Understanding Data Similarity in Large-Scale Scientific Datasets

Payton Linton, William Melodia,
Alina Lazar
Youngstown State University
Youngstown, USA
palinton, wmmelodia@student.ysu.edu,
alazar@ysu.edu

Deborah Agarwal, Ludovico Bianchi,
Devarshi Ghoshal, Gilberto Pastorello,
Lavanya Ramakrishnan, Kesheng Wu
Lawrence Berkeley National Laboratory
Berkeley, USA
daagarwal, lbianchi, dghoshal@lbl.gov
kwu, gzpastorello, lramakrishnan@lbl.gov

***Abstract*—Today, scientific experiments and simulations produce massive amounts of heterogeneous data that need to be stored and analyzed. Given that these large datasets are stored in many files, formats and locations, how can scientists find relevant data, duplicates or similarities? In this context, we concentrate on developing algorithms to compare similarity of time series for the purpose of search, classification and clustering. For example, generating accurate patterns from climate related time series is important not only for building models for weather forecasting and climate prediction, but also for modeling and predicting the cycle of carbon, water, and energy. We developed the methodology and ran an exploratory analysis of climatic and ecosystem variables from the FLUXNET2015 dataset. The proposed combination of similarity metrics, nonlinear dimension reduction, clustering methods and validity measures for time series data has never been applied to unlabeled datasets before, and provides a process that can be easily extended to other scientific time series data. The dimensionality reduction step provides a good way to identify the optimum number of clusters, detect outliers and assign initial labels to the time series data. We evaluated multiple similarity metrics, in terms of the internal cluster validity for driver as well as response variables. While the best metric often depends on a number of factor, the Euclidean distance seems to perform well for most variables and also in terms of computational expense.**

***Index Terms*—dimensionality reduction, clustering, similarity measure**

I. Introduction

Currently, there are 2.5 quintillion bytes of electronic data [1] created every day and this pace is not going to decrease in the near future. Scientific data collected from scientific observations, experiments, and large-scale simulations with provenance in different scientific domains, such as earth and space science, astronomy, genomics, environment, and physics, follow the same trend [2]. The size of these datasets today typically ranges from hundreds of gigabytes to tens of petabytes. For example, in 2017, the data collected from the Large Hadron Collider (LHC) has passed 200 petabytes [3]. A number of Department of Energy's (DOE) applied science offices generate extensive experimental and observational scientific data that require computational, networking and storage resources for processing, transfer and analysis. This data is prone to frequent changes and updates due to changes in instrument configurations, software updates or data cleaning. However, scientists often do not have enough information about these data changes to track them efficiently and to make decisions about their impact on the actual data processing and analysis [4]. Currently, there is no systematic and organized way to collect and track information about the types of changes and the amount of data change [5]. Researchers often need to re-run entire data processing pipelines when updated datasets are released. As a part of the Deduce project[1] at Lawrence Berkeley National Laboratory (LBNL), a team of computer scientists are investigating an end-to-end methodology in addition to building software tools [6] for identifying, capturing and tracking data changes in large scientific data collections.

The first step in building robust methods for tracking data changes is to reformulate the problem as a similarity search. In general, this depends on the type of data, but for time series data, similarity search is defined as the query operation that finds the set of data series closest to a given series. The results returned by this query are ranked according to some definition of distance or similarity. A data series similarity metric is a function that measures the (dis)similarity of two data series and it is instrumental in measuring change. Many similarity metrics have been proposed in the literature [7], especially for classification, however the Euclidean distance remains one of the simplest and the most widely used, as well as one of the most effective for large data series collections in terms of computation cost.

In this paper, we evaluate and compare similarity measures for time series for the purpose of similarity search and clustering. For example, to understand climate models and especially land surface models, we need to first analyze the temporal and spatial similarity or variability for driver variables and certain atmospheric conditions (such as exchange of heat, moisture and various carbon fluxes, etc.). Well-known algorithms, including dimensionality reduction and clustering, might be effective for these tasks, however, a number of challenges in this type of data make it difficult to even perform the basic comparison operations. One core challenge is that the time series data is high dimensional and there is no perfect solution in terms of the metric used to compute

[1] http://deduce.lbl.gov

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

similarity. Moreover, to compare data collected from two sites, we have to compare a number of different types of attributes. Additionally, the time series frequently contain missing values, unknown values or corrupted values, which presents other challenges. Our aim is to investigate how different time series similarity measures affect state-of-art dimensionality reduction and clustering techniques.

The rest of the paper is organized as follows. We present related work in section II. We describe the machine learning methods in section III and the data in section IV. We present our results in section V and conclusions in section VI.

II. Related Work

At the heart of this research lies problems relating to time series similarity, correlation, clustering and prediction. These problems are by no means new and can be seen throughout many different applications in several areas of study.

A. Time Series Similarity

Time series similarity measures provide a quantitative comparison that can be applied to time series to capture amplitude timing and noise effects. One such study by Wang et al. [8] looked into nine different similarity metrics applied to a variety of time series datasets. The study concluded that not all similarity metrics are equal in terms of efficiency and accuracy and that different similarity metrics will preform better on different datasets based on a multitude of different properties within the dataset.

Another similar study by Serr et al. [9] looked at seven of the most prominent similarity metrics in depth and applied them to several time series datasets with different properties. The conclusion was that there are several similarity metrics that are benifical for a variety of applications within time series data and several that fell short despite being promising candidates.

These studies can be further applied to many different problems, including detecting changes and similarities in ecosystems [10] and climate environments. The FLUXNET2015 dataset [11] is often used for meteorological and environmental research studies to answer questions about temporal variability or to prove correlation and causality between variables. The majority of these studies are done at a small scale and usually use only statistical methods for analysis.

B. Statistical Approaches

Recently, Baldocchi et al. [12] analyzed data from 59 sites over five or more years using the classical standard deviation to compute the inter-annual variability in net ecosystem carbon exchange to show carbon fluxes changes over the years.

Chu et al. [13] focused on the temporal representatives of FLUXNET sites to determine if the set of measurements collected at any given site can capture the natural variability of climatological driver conditions and extrapolate or predict unknown and/or future measurements. They found that the temporal similarity of driver variables provides good results for the response variables' extrapolation.

To predict site behavior across the network, data from 155 sites were used for the analysis presented in [14]. They quantify predictability of the key fluxes in terms of site uniqueness. The analysis is based on clustering combined with multiple linear regression models that fit functions between the driver and the response variables. By doing this, they were able to conclude that the drier the site, the more unique it is. However, based on their definition of uniqueness, data length, quality vegetation type had little impact on the uniqueness of the site.

Cui et al. [15] summarized the energy balance closure (EBC) data from site residing in nine different vegetation zones and five climate. They also explored the correlation between the driver variables and EBC. Their results showed that EBC is closely related with air temperature, precipitation, friction velocity, vapor pressure deficit (VPD) and enhanced vegetation index.

C. Machine Learning Approaches

Machine learning has been an important tool for understanding carbon fluxes. For example, Murphy et al. [16] applied multiple machine learning algorithms on FLUXNET data to predict the carbon flux. They evaluate the performance of four classes of machine learning: artificial neural networks, Gaussian process regressions, random forests, and recurrent neural networks implemented using long short-term memory. They also identified input variables that contribute most to the carbon flux predictions.

However, Reichstein et al. [17] claimed that current approaches may not be optimal when system behavior is dominated by spatial or temporal context. They claim that using deep learning on these contextual cues would be better to further the understanding of Earth system science problems.

Our approach is similar to some of the previous studies that research temporal and spatial variability in the FLUXNET data. We use dimensionality reduction combined with clustering to evaluate several similarity measures. The clustering is performed on data aggregated by day. The approach has been applied to the combination of driver variables and also to each response variable individually.

III. Methods

Under the assumption that these FLUXNET sites fall into a few categories despite the wide geographic and ecosystem differences they have, we executed a clustering method to extract patterns from the multivariable time series data. We want to not only identify such groups, but also find trends within one or more variables [18].

Our process uses a multi-step approach. First, similarities containing a quantification of closeness is determined for each pair in the dataset. This is done using multiple similarity measures. Next, a combination of clustering algorithms and different numbers of clusters are evaluated by multiple clustering validity measures. Lastly, the effects of the different similarity measures on the clustering solutions are examined both quantitatively and qualitatively.

A. Dimensionality Reduction

For dimensionality reduction and visualization, we employ the newly developed Uniform Manifold Approximation and Projection (UMAP) [19] algorithm. Unlike Principal Component Analysis (PCA) that captures linear trends in the data, UMAP is a dimensionality reduction method that works well for embedding nonlinear data into a low dimensional space for visualization. UMAP framework is based in a Riemannian geometry and algebraic geometry. This algorithm is based on the following assumptions:

- the data is uniformly distributed on a Riemannian manifold
- the Riemannian metric is locally constant
- the manifold is locally connected

The resulting manifold is then modeled with a fuzzy topological structure. Finally, the low dimensional embeddings are found by a similarity search for a projection of the data on the fuzzy topological structure.

The proposed methodology can be done without performing the dimensionality reduction step, however especially for high dimension, multivariate data this step provides a 2-dimensional representation of the data. This representation can be used to visualize the underlying structures may exist in the data and also may improve the clustering results.

B. Similarity Metrics

Since similarity measures such as the Euclidean distance, may not be ideal for computing similarities between large time series datasets, alternative similarity metrics need to be explored [7]. Such measures are generally used for machine learning algorithms such as classification and clustering [20], [21]. Four similarity metrics are considered for the purpose of this study: Euclidean Distance, Fourier Coefficients, Simple Correlation, and Dynamic Time Warping (DTW).

1) Euclidean Distance: The Euclidean distance metric is the most common method for determining the similarity in terms of the distance between any two objects or data points [22]. The Euclidean distance metric uses the Pythagorean Theorem to calculate the distance between two given points. As one might expect, the smaller the output of the distance, the more similar the two points are. This algorithm works with multi-dimensional data, making it a good candidate for this high dimensional meteorological data. However, the Euclidean distance focuses solely on linear distance, meaning any treads throughout time series will be ignored. For massive datasets, Euclidean distance can be computed very fast giving this method a definite advantage over other methods that are quite computationally expensive.

2) Fourier Coefficients: Another common similarity metric is called Fourier coefficients. This approach transforms the incoming time series into Fourier space and then measuring their Euclidean distance in Fourier space. Because the Fourier transforms requires time to compute, this similarity metric typically requires more time to compute than computing the Euclidean distance. It is possible that the Fourier transforms are available, then the computational cost of this approach could be actually lower, especially when only some of the lower frequency components are used for the similarity calculation. Of course, if only the lower frequency components are used, this similarity metric might not capture the high-frequency patterns in the initial data.

3) Simple Correlation: An outside the box approach that we consider is Simple correlation. Simple correlation treats the data as separate vectors and computes the statistical correlation between the vectors [23]. This approach is more of a measure of dependence rather than a traditional distance, but it works for our purposes. The result of this measure is somewhere between 0 and 1, where 0 is completely independent with no relation between the vectors at all and 1 is complete dependence, meaning the vectors are exactly the same.

4) Dynamic Time Warping: Another method to determine the similarity between two time series is Dynamic Time Warping (DTW). This method take into account the time shift between the two time series, as shown in Figure 1. This is an improvement over other methods because DTW can cluster through time as well as space [24]. It creates an optimal match between two sequences regardless of any time differences. One trade off of this flexibility in computing the similarity is that for n terms, the algorithm has to make n^2 calculations, making this method computationally expensive for large datasets.

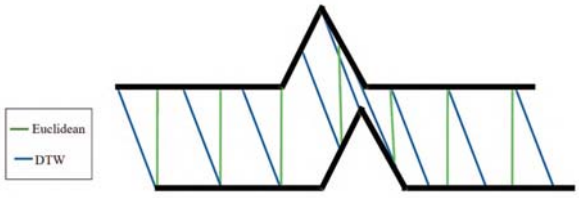


Fig. 1: Example of DTW vs Euclidean, where the two black lines are datatsets and the colored lines are the comparisons made by each method.

To extend this idea to multi-channel or multivariate time series or sequences, the new similarity metrics have to include the contribution of each given variable in the final computation. To accomplish this, the distances for each individual variable are first computed using one of the regular distance measures. Then, the multivariate similarity is calculated by averaging the values for the results from the previous step.

C. K-Means Clustering

Once a similarity matrix for all pairs of data points has been computed, a clustering approach can be used to group data points based on their similarity. K-Means [25] is a well-known unsupervised clustering method that works both well and quickly. The algorithm partitions the data points into k groups. At the initialization step, the algorithm will randomly chooses k data points as the clusters' centers. The data points are then distributed to the clusters based on their distances

to the clusters' centers. The algorithm updates and refines the clusters by calculating the distances of each data point to the clusters' centers. The data points are then placed into the cluster with the smallest distance. The centers of the clusters are recalculated at each iteration until the algorithm converges to the solution and the refinement process no longer makes any changes. The algorithm usually only requires a few iterations to converge.

D. Internal Cluster Validity Measures

To evaluate clustering methods on datasets such as the FLUXNET2015 time series data, for which there is no "ground truth", accuracy and error do not work. Alternative approaches for evaluation of the clustering algorithms must be implemented and used. Several internal clustering validity measures [26], [27] have been proposed to provide a statistical quality measure for the generated partitions. These internal measures include the Calinski and Harabasz score (CHS), the Davies-Bouldin score (DBS), and the Average Silhouette Width (ASW).

The CHS measure is defined as the average between the within-cluster sum of squares and the between-cluster sum of squares. The DBS measure is the ratio between the average distances within of one cluster and the in between cluster distances. This results in a better score for clusters which are farther apart and less dispersed. The ASW is calculated for each data point as the average distance to the other points in its cluster and compared with the average distance to the points in the next closest cluster.

These measures establish a way to choose the optimal number of clusters without relying on the ground truth. The results of these measures are represented in a tabular or in a plot, allowing for visual interpretation. Once the best clustering method and the optimal number of clusters have been selected, these parameters are used to generate the clustering groups. The best way to identify general cluster characteristics for sets of sequences is by plotting the mean, min, and max of the time series that belong to the same cluster.

IV. Datasets

To determine the effectiveness of these methods, this research uses the FLUXNET2015 dataset [11], which includes measurements of the exchange of carbon, water, and energy between terrestrial ecosystems and the atmosphere using eddy covariance method, which is the largest of its kind [13]. The dataset includes 212 sites from all continents except Antarctica, with over 1500 site-years of data collected at 30 or 60 minute resolution and distributed at multiple aggregated resolutions (hourly, daily, monthly, etc.) The dataset is collected by multiple regional networks globally and distributed using the Fluxdata.org[2] platform hosted by the Lawrence Berkeley National Laboratory. A selection of FLUXNET sites is shown in Figure 2.

The data is run through quality control to ensure accuracy and remove faulty data. However, many of the variables

[2]https://fluxnet.fluxdata.org/

Fig. 2: Location of clusters when looking at air temperature, shortwave radiation, precipitation, VPD and wind speed where the diamonds correspond to the green cluster, the triangles correspond to the black cluster, the squares correspond to the yellow cluster, and the circles correspond to the blue cluster.

contain missing data, mostly due to power issues or faulty instruments giving inaccurate readings [28]. To retrieve these meteorological variables outdoor instuments are used, and as such, are exposed to the varying weather patterns and natural disasters that may occur in an area.

Data can also be missing for years based on when the tower began or stopped collecting data. Though this study is focused specifically on the FLUXNET2015 dataset, it hopes to develop general methods that would work on any datasets. The variables considered for this project are the driver variables – temperature, shortwave radiation, vapor pressure deficit (VPD), wind speed, and precipitation. Where shortwave radiation is the incoming radiation from the sun and vapor pressure deficit is the difference between the amount of moisture in the air and the amount of moisture that the air can hold a full saturation. All these variables impact the evaluated response variables: sensible heat, latent heat, and two different processed net ecosystem exchange values. The sensible heat is a measure the heat absorbed and released into the air with changing temperatures, while the sensible heat is a measure of the heat absorbed an released into the atmosphere when the air is subject to a phase change, such as precipitation or condensation. The Net Ecosystem Exchange is a measure of the total carbon transfered between the earth and the atmosphere. These driver and response variables were chosen for their completeness – they are gap-filled as part of the processing done for the FLUXNET2015 dataset.

V. Experiments and Results

The proposed methodology can approximately identify near-duplicate datasets from massive collections of datasets by computing small representations for each dataset and comparing only these reduced sets and not the entire datasets, for fast performance and without loss of quality. Looking at the response variables individually, the best similarity metrics varied. For latent heat and sensible heat, the correlation distance performed better. However for temperature and both

TABLE I: Internal Clustering Validity Measures for Driver Variable

	ASW	CHS	DBS
Correlation	0.54297	**16248.8**	0.56670
DTW	0.60120	15763.4	0.42803
Euclidean	**0.67947**	9262.15	**0.32295**
Fourier	0.67098	9958.30	0.33138

variations of the net ecosystem exchange, DTW provided the best performance. These results Table I used cluster validity measures and analyzed the values that these measures gave. Although with the driver variables, when the variables were considered simultaneously, the cluster validity gave different results, showing the Euclidean distance was best as shown in Table I.

TABLE II: Internal Clustering Validity Measures for Response Variable

	Euclidean	Fourier	Correlation	DTW
Sensible Heat				
ASW	0.525077	0.449927	**0.675231**	0.463661
CHS	3241.57	2103.77	**5456.11**	2123.26
DBS	0.852836	0.887615	**0.699878**	0.894706
Net Ecosystem Exchange (VUT_REF)				
ASW	0.530670	0.499239	0.479826	**0.566473**
CHS	2629.77	2572.49	2343.05	**3166.83**
DBS	0.892706	0.942261	0.966203	**0.868652**
Latent Heat				
ASW	0.488126	0.465864	**0.636779**	0.602823
CHS	2459.43	2038.08	**6726.95**	3759.65
DBS	1.05667	1.06461	**0.659145**	0.845184
Net Ecosystem Exchange (CUT_USTAR50)				
ASW	0.518020	0.516414	0.475644	**0.558803**
CHS	2819.97	2725.23	28358.25	**3038.71**
DBS	0.929540	0.964514	0.919367	**0.890048**

A. Driver Variables Clustering

When considering the combination of driver variables (air temperature, precipitation, shortwave radiation, VPD and wind speed) we found that the Euclidean distance algorithm was the best (Table I. Thus, we used that algorithm and clustered the data using K-means for multiple clusters ranging from two to ten. Next, cluster validity metrics, including the Average Silhouette Width(ASW), the Calinski Harabasz Score(CHS), and the Davies Bouldin Score(DBS), were run for each number of clusters until the best number of clusters was found. The optimal number of clusters occurs when the Average Silhouette Width and the Calinski Harabasz Score have a peak and when the Davies Bouldin Score is at a local minimum. As seen in Fig. 4, the ideal number of clusters is four. The Silhouette plot in Fig. 5 proves the cluster solution is valid, since the silhouette coefficients of all the clusters are higher than the average value.

B. Driver Variables Cluster Interpretation

Checking the cluster distribution in terms of the latitude and longitude of the site (Figure 2 and the year the data was recorded, we know specific information about every cluster. The yellow cluster contains only sites near the US-Mexico border, at about 30°N. The black cluster mainly consists of sites between 35°and 50°N. The green cluster consists of sites mainly above 50 °N, but also include some sites at high elevations. The main difference between the green and black cluster is that the peak air temperature is lower for the sites in the green cluster. The blue cluster contains sites mainly between 0°and 35°S, with the majority of the sites being in Australia.

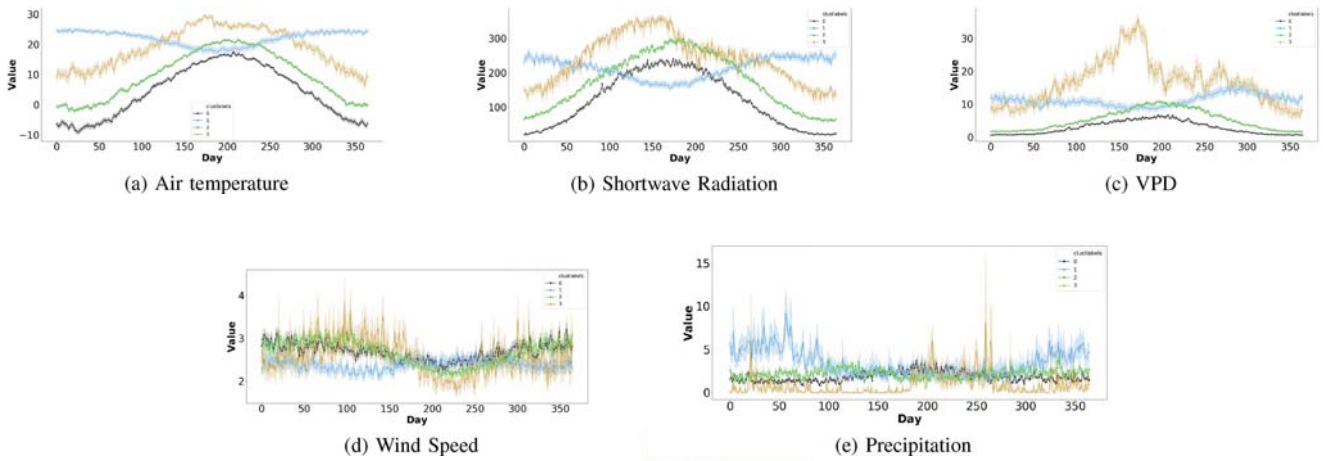
C. Driver Variables Cluster Shape Representation

Aggregates over the time series for each cluster and each of the driver variables are shown in Figure 3. For air temperature, shortwave radiation, and VPD, the yellow cluster has the highest peak overall and always has higher values than the green and black clusters, while the blue cluster follows the opposite patterns as the others due to the seasons being opposite in the southern hemisphere. Fig 2 shows the locations of all clusters where the colors match the cluster color. For precipitation, we see the the yellow and blue clusters get precipitation in waves while the black and green have consistent precipitation year round. For wind speed, all of the clusters follow about the same trend except the blue cluster, which has a unique trend. These variables are dependent on the season and latitude primarily, but are also dependent on the vegetation type and altitude to a lesser degree.

D. Analysis for the Response Variables

The use of cluster validity showed that sensible heat and both of the net ecosystem exchange variables were best split into five clusters (Figure 8), while the latent heat was placed into seven clusters (Figure 7). To look at how the sites were allocated to the clusters, the data was used to make shape profiles shown in Figure 6. The shape profiles were created based on the which distance algorithm best suited the variable, Table II shows the values of the cluster validity metrics for the different response variables that were analyzed. According to these values correlation best fits the heat fluxes while DTW provides the best measures for the different NEE values. These distance metrics were then used to create the corresponding shape profiles. For the heat variables, many of the sites were split based on the hemisphere that the sites were located in. Clusters that are concave up are composed of many sites in the southern hemisphere, while the concave down plots are sites located in the northern hemisphere. The shape profiles show the line of the average value calculated for each day in the year, while the shaded region shows the range of values that were observed on that day.

The clusters made from the net ecosystem exchange values do not appear to cluster based on geographical region. It is currently unknown how these sites are being clustered. However, initial analysis suggests that the sites are being clustered based on vegetation grown in the sites area,or possibly the altitude of the site has an affect on the NEE variables.



(a) Air temperature (b) Shortwave Radiation (c) VPD

(d) Wind Speed (e) Precipitation

Fig. 3: All driver variable shape profiles

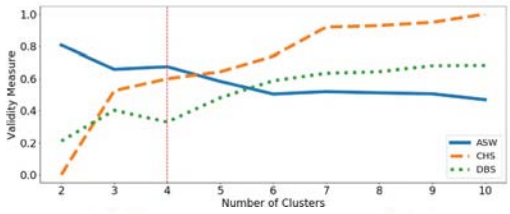


Fig. 4: Result of cluster validity

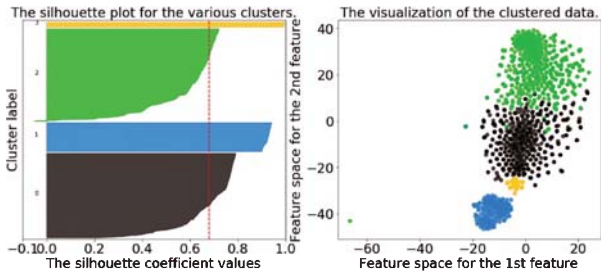


Fig. 5: Silhouette plot for combined driver variables

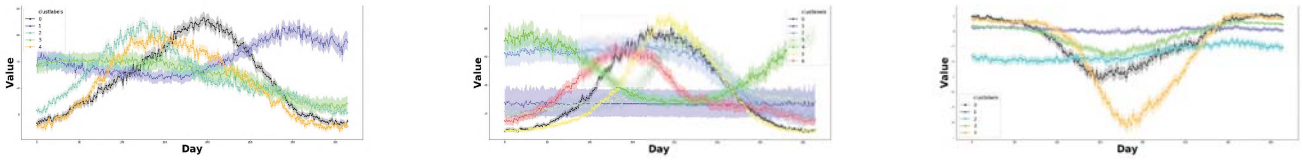
VI. Conclusions

Using clustering and dimensionality reduction to analyze the results and performance of similarity metrics, we detected temporal and spatial similarities in the FLUXNET2015 dataset. After testing multiple similarity algorithms using Euclidean distance, DTW, simple correlation and Fourier coefficients, we concluded that Euclidean distance gave the best combined values for the cluster validity metrics when run with the driver variables dataset. However, this did not hold true for the response variables. For the heat variables simple correlation offers the best results; however, Dynamic Time Warping works best with the net ecosystem exchange variables. With this information, we decided to continue mainly with the Euclidean distance since the minor additional benefits received from Dynamic Time Warping is not worth the additional computational time that is required in the acquisition of the marginally better measurements. The development of these similarity algorithms will help researchers be able to choose appropriate similarity metrics to get similarity results that best represent the data being considered.

VII. Future Work

We plan to extend this research to include all the relevant variables in the dataset in the analysis. We would like to determine the most efficient similarity measure when using the all variables and potentially run a feature selection-type of analysis. This would be helpful in finding thoroughly evaluated similarity measures within the larger dataset. Many scientists are interested on the correlation between the driver and response variables, and the similarity measures used in this research can be help identify these relationships and bring to light new scientific knowledge still locked within these rich and complex types of datasets.

Acknowledgments. This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by Office of Workforce Development for Teachers and Scientists (WDTS) under the Visiting Faculty Program (VFP), Office of Science, the U.S. Department of Energy. This research used resources of the National Energy Research Scientific Computing Center. This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices.



(a) Sensible Heat (b) Latent Heat (c) Net Ecosystem Exchange

Fig. 6: All Response Variable shape profiles

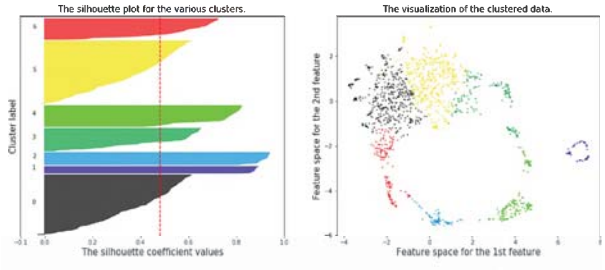


Fig. 7: Silhouette plot for Latent Heat response variable

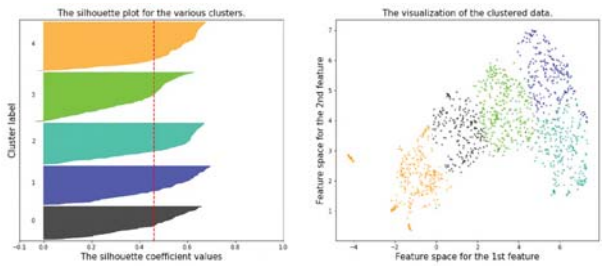


Fig. 8: Silhouette plot for Net Ecosystem Exchange response variable

REFERENCES

[1] (2017, June) Data never sleeps 5.05. [Online]. Available: https://www.domo.com/learn/data-never-sleeps-5

[2] C. A. Mattmann, "Computing: A vision for data science," *Nature*, vol. 493, no. 7433, pp. 473–475, Jan. 2013.

[3] M. Gaillard. (2017, July) Cern data centre passes the 200-petabyte milestone. [Online]. Available: https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone

[4] D. Shanmugam, D. Blalock, and J. Guttag, "Jiffy: A convolutional approach to learning time series similarity," 2018.

[5] A. Z. Broder, "Identifying and filtering near-duplicate documents," in *Annual Symposium on Combinatorial Pattern Matching*. Springer, 2000, pp. 1–10.

[6] D. Ghoshal, L. Ramakrishnan, and D. Agarwal, "Dac-Man: Data change management for scientific datasets on HPC systems," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, ser. SC '18. Piscataway, NJ, USA: IEEE Press, 2018, pp. 72:1–72:13.

[7] J. Serra and J. L. Arcos, "An empirical evaluation of similarity measures for time series classification," *Knowledge-Based Systems*, vol. 67, pp. 305–314, 2014.

[8] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 275–309, 2013.

[9] J. Serrà and J. L. Arcos, "An empirical evaluation of similarity measures for time series classification," Jan. 2014.

[10] S. Lhermitte, J. Verbesselt, W. W. Verstraeten, and P. Coppin, "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics," *Remote sensing of environment*, vol. 115, no. 12, pp. 3129–3152, 2011.

[11] G. Pastorello, D. Papale, H. Chu, C. Trotta, D. Agarwal, E. Canfora, D. Baldocchi, and M. Torn, "A new data set to keep a sharper eye on land-air exchanges," *Eos, Transactions American Geophysical Union*, vol. 98, no. 8, 2017.

[12] D. Baldocchi, H. Chu, and M. Reichstein, "Inter-annual variability of net and gross ecosystem carbon fluxes: A review," *Agricultural and Forest Meteorology*, vol. 249, pp. 520–533, 2018.

[13] H. Chu, D. D. Baldocchi, R. John, S. Wolf, and M. Reichstein, "Fluxes all of the time? a primer on the temporal representativeness of fluxnet," *Journal of Geophysical Research: Biogeosciences*, vol. 122, no. 2, pp. 289–307, 2017.

[14] N. Haughton, G. Abramowitz, M. G. D. Kauwe, and A. J. Pitman, "Does predictability of fluxes vary between fluxnet sites?" *Biogeosciences*, vol. 15, no. 14, pp. 4495–4513, 2018.

[15] W. Cui and T. F. M. Chui, "Temporal and spatial variations of energy balance closure across fluxnet research sites," *Agricultural and Forest Meteorology*, vol. 271, pp. 12–21, 2019.

[16] D. Murphy, B. Bonner, G. S. Nearing, C. Pelissier, and M. Halem, "Machine learning for carbon monitoring," in *AGU Fall Meeting Abstracts*, 2018.

[17] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, p. 195, 2019.

[18] T.-c. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011.

[19] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," Feb. 2018.

[20] 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, no. 4, pp. 349–371, 2003.

[21] T. Santos and R. Kern, "A literature survey of early time series classification and deep learning." in *Sami@ iknow*, 2016.

[22] A. Abanda, U. Mori, and J. A. Lozano, "A review on distance based time series classification," *Data Mining and Knowledge Discovery*, vol. 33, no. 2, pp. 378–412, 2019.

[23] G. J. Székely, M. L. Rizzo, N. K. Bakirov *et al.*, "Measuring and testing dependence by correlation of distances," *The annals of statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[24] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[25] T. Thinsungnoena, N. Kaoungkub, P. Durongdumronchaib, K. Kerdprasopb, and N. Kerdprasopb, "The clustering validity with silhouette and sum of squared errors," *learning*, vol. 3, p. 7, 2015.

[26] S. A. L. Mary, A. Sivagami, and M. U. Rani, "Cluster validity measures dynamic clustering algorithms," *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 9, 2015.

[27] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 911–916.

[28] D. Hui, S. Wan, B. Su, G. Katul, R. Monson, and Y. Luo, "Gap-filling missing data in eddy covariance measurements using multiple imputation (mi) for annual estimations," *Agricultural and Forest Meteorology*, vol. 121, no. 1-2, pp. 93–111, 2004.