# IC² Faculty Research Program

## Final Report

# A Machine Learning Approach: Socio-economic Analysis to Support and Identify Resilient Analog Communities in Texas.


**By**

Ademide O. Mabadeje


**IC² Institute Final Report**


**Supervised by:**

Dr. Michael Pyrcz (Advisor)

**The University of Texas at Austin**


August 26, 2022

## Abstract

Identification of analog resources or items are important during the planning and development of new communities because available information is usually limited or absent. Conventionally, analogs are made by domain experts however, this is not always readily obtainable. Coupled with this challenge, most of the available data in socioeconomic systems have high dimensionality making interpretation, and visualization of these datasets difficult. Hence, it is crucial to adopt a workflow that can be used to identify analogs regardless of its existing high dimensionality.

To this end, we present a systematic and unbiased measure, group similarity score (GCS) and similarity scoring metric (SSM) to support the predictive search of missing properties for target communities and identification of analogous cities based on available socioeconomic data and modeling. Knowing that each Texan community can be characterized by its associated properties, the workflow combines both spatial and multivariate statistics in a novel manner to determine the GCS & SSM whilst visualizing the associated uncertainty space.

The workflow consists of three major steps: 1) key parameter selection via feature engineering, 2) multivariate and spatial analysis using multidimensional scaling (MDS) and density-based spatial clustering of applications with noise (DBSCAN) for clustering analysis, 3) similarity ranking using a modified Mahalanobis distance function as a clustering basis on preprocessed data. Afterwards, to assess the quality of the predicted feature and analog communities obtained, K-nearest neighbor algorithm is applied, then the analog cities are found.

The workflow is demonstrated using on high dimensional socio-economic data. We find analogs for each community cluster identified with their GCS and SSM in relation to 4 randomly selected communities used for testing. Thus, it is recommended to apply the integration of this workflow in uncertainty exploration, trend-mappings, and community analog assignment, and benchmarking to support decision making.

## 1. Introduction

This section reviews previous and current methods used for analog assignment and selection from a petroleum engineering standpoint. The appropriate use of analogs is an important factor during resource assessment and reserve estimation in oil and gas (Hodgin & Harrell, 2006;

Sidle & Lee, 2010). Predominantly, analogs are used to assess economic producibility, production decline characteristics, drainage areas, and total recoverable resource. Also, analogs are imperative in every asset management team to leverage insights that can maximize hydrocarbon recovery and inform development decisions (Smalley et al., 2009; Gomes et al., 2018, Masoudi et al., 2020). Albeit a crucial component, most analog analyses are carried out in a qualitative manner solely dependent on human expertise, experience, and conventional statistical methods. This results in subjective workflows riddled with bias, lack of repeatability, and inhibited insights from best practices and past mistakes.

Over the past 25 years, there have been many attempts to develop quantitative analog indexing methods for reservoirs using a wide range of reservoir and geological input. Dromgoole and Speers (1997) spearheaded attempts to describe reservoir complexity using a field scoring method for UK North Sea fields. The scoring method was only based on "geological complexity" and input parameters requiring high-level interpretation, resulting in a biased and irreplicable application. A more structured method, reservoir complexity index (RCI) was introduced by Bygdevoll (2007) using parameters assigned based on a combination of objective limits and subjective assessment. Although objectivity is infused, RCI has the same limitations as Dromgoole and Speers (1997) because non-repeatable assessments are required, and the criteria are specific to the basin of interest, therefore limiting generalization. Sun & Pollitt (2021) developed a 5-step heuristic approach for analog quantification indexes in reservoirs. As the heuristic approach addresses the repeatability issue, Simpson's paradox of mixing populations is introduced by creating a global analog in highly heterogeneous reservoirs. Also, subjectivity related to parameter selection was reduced by using fine sequential filter selection on geologic, fluidic, and engineering properties, however, it was not completely mitigated because of its rubric- based parameter selection.

With the 4$^{th}$ paradigm at its peak, data-driven algorithms and similarity measures have been combined by researchers to identify analog reservoirs. As a result, Bhushan et al. (2002) identifies reservoir analogs using the smart reservoir prospector (SRP) – a metric using the nearest neighbor algorithm to measure the degree of similarity between reservoirs weighted by each reservoir attribute. Although a significant step towards statistical analog identification, SRP is flawed because it neglects the distortion of distance in a high dimensional space by computing similarity from Euclidean distance. In Rodriguez et al. (2013), the effect of distance distortion prompted the use of principal components analysis (PCA) for dimensionality reduction and co-linearity avoidance between reservoir properties. Then, Ward's hierarchical clustering was used to generate a similarity ranking of analogous reservoirs (Rodriguez et al., 2013). Olukoga & Feng (2021) applied heuristic algorithms – k-means, k-medians – and hierarchical clustering algorithms to find miscible $CO_2$ flooding analogous projects. The authors performed PCA for dimensionality reduction and used a principal component-weighted Euclidean distance as a similarity measure using K-means.

However, a challenge with such heuristic algorithms is that the number of classes (clusters) to be determined is assigned. Moreover, PCA assumes a linear relationship between the data and underlying latent variables represented as principal components. Meanwhile, the relationships

between most variable types in unconventional plays are non- linear, hence, the coinage statistical plays. The preceding statements validate the need for an alternative clustering algorithm and dimensionality reduction method during analog identification that accounts for spatial settings. Currently, there is no known objective metric for summarizing high dimensional cases of features that group and identify geological analog wells while also accounting for their spatial settings.

## 2. Methodology

To achieve the project's objectives, the following phases and corresponding steps are followed:

1. *Dissimilarity matrix calculation*
   Distances are distorted in high-dimensional spaces (Köppen, 2000). Therefore, there is a need for a distance metric that circumvents the distortion given that Euclidean and Manhattan distances are not applicable. Hence, Mahalanobis distance is effective because it performs well with multivariate, highly dimensional, correlated data. However, to introduce a physics-based constraint, a novel weighted Mahalanobis distance using scaled mutual information is integrated to account for the systems' spatial settings. Then, the dissimilarity matrix is sorted using hierarchical clustering with Ward linkage as a diagnostic plot to check if there are inherent clusters in the matrix.

2. *Multidimensional Scaling (MDS)*
   MDS is a nonlinear dimensionality reduction technique used to preserve a measure of similarity or dissimilarity between pairs of data points by projecting multidimensional data into a lower dimensional space (Kruskal, 1964; Cox & Cox, 1994). Knowing that non-metric MDS applies to ordinal data, and that classical MDS is simply PCA with a Euclidean distance, metric-MDS is chosen as the dimension reduction method. Scheidt & Caers (2009) found that MDS retains intrinsic information and spatial context of the data. Further, Tan et al. (2014) showed MDS can also be used as a measure of uncertainty space for various spatial models. MDS uses the previously determined dissimilarity matrix as input and then gives 2D projections.

3. *Ordinary kriging on features in original feature space and projection space*
   The kriged estimate and variance of the response feature are computed in both Euclidean feature space and MDS space. Next, the kriged response features in both spaces are used as underlying spatial maps to identify trends if any.

4. *Clustering analysis to identify community/ city groupings/ labels*
   Following the derivation by Ester et al. (1996) and guidelines from Schubert et al. (2017) on density-based spatial clustering of applications with noise (DBSCAN), the MDS projections are clustered to identify groupings in the subspace. To ensure the optimal number of clusters will be found via DBSCAN, hyperparameters– "min. pts" and "eps"–will be tuned. Where "min. pts" is the minimum number of samples required to form a cluster, and "eps" is the maximum radius of the neighborhood used for

expanding clusters. Then, the DBSCAN algorithm will be fitted using the tuned hyperparameters.

5. *Classification-based predictive modeling*
   Since the main research objective directly focuses on finding similarities between communities, the K-nearest neighbor classifier (KNN) is chosen because of its inherent nature to perform local optimizations. For this workflow, KNN is a better classifier as it does not assume independence between features when compared to the Naïve Bayes classifier. The KNN classifier is implemented, and the cluster groups found by DBSCAN are used for prediction purposes. Next, the optimal K-

value is determined via a k-crossfold validation on the training set using inverse- distance weighting. Then, the now-trained model using prior parameters is used to predict the cluster grouping in the test dataset. To check model goodness, a classification report consisting of accuracy, precision, and F1-score is generated alongside the confusion matrix to determine the fraction of misclassified labels and otherwise. Lastly, a probabilistic uncertainty scheme will be used to identify communities in a unique cluster, and boundary communities with tendencies of belonging to more than one cluster to account for grouping uncertainties.

6. *Analog identification and similarity scoring*
The centroid of each cluster identified in the MDS space is determined as an analog. Then a within-group and between-group similarity score called similarity scoring metric (SSM) and group consistency score (GCS) is computed in the normalized MDS space. These scores were developed using Euclidean distance and can be ranked in ascending order with the most similar communities ranked first; where 0 indicates complete similarity, and 1 indicates total dissimilarity.

## 4. Results & Discussion

This section discusses the results obtained from the research objectives. These outcomes particularly address multivariate spatiotemporal analysis and machine learning to support analog community studies in Texas. The dataset needed to achieve our objectives should include appropriate sociological, economic, and census features of importance after the implementation of feature engineering on a normalized scale. Due to the unavailability of data at IC², a synthetic but realistic dataset comprising 4 socio-economic factors: median household income, population, crime rates, and commute time was created using appropriate relationships and correlations as found in the literature. Where median household income is the response feature, and the remaining factors are predictor features.

Identification of analog items or resources is important because available information about new areas is usually limited or non-existent. Traditionally the search for analogs is done by experienced domain experts, but this practice is subject to the availability of this experience and the results are heavily dominated by preferences. From a petroleum engineering standpoint, most methods in the literature for analog

identification and classification are either applied to reservoirs or rock facies with no quantitative analog indexing approach using geological input data for wells. Consequently, the idea of using knowledge from well-known comparable reservoirs with fluid and reservoir properties identical to an undeveloped target reservoir is extended to "geological well typing". Not only is geological well typing an intrinsically complicated multivariate spatiotemporal problem, but its challenges are also exacerbated due to its high dimensional nature (Mabadeje & Pyrcz, 2022b).

Here, a dataset with high dimensionality is generated using adequate relationships and correlations inherent in literature for the following socio-economic factors on a city level: The response feature: household income ($K), and the predictor features: crime rates (%), population (K), and commute time to work (minutes). To understand the data, a bivariate matrix scatter plot is shown in Figure 4–1.

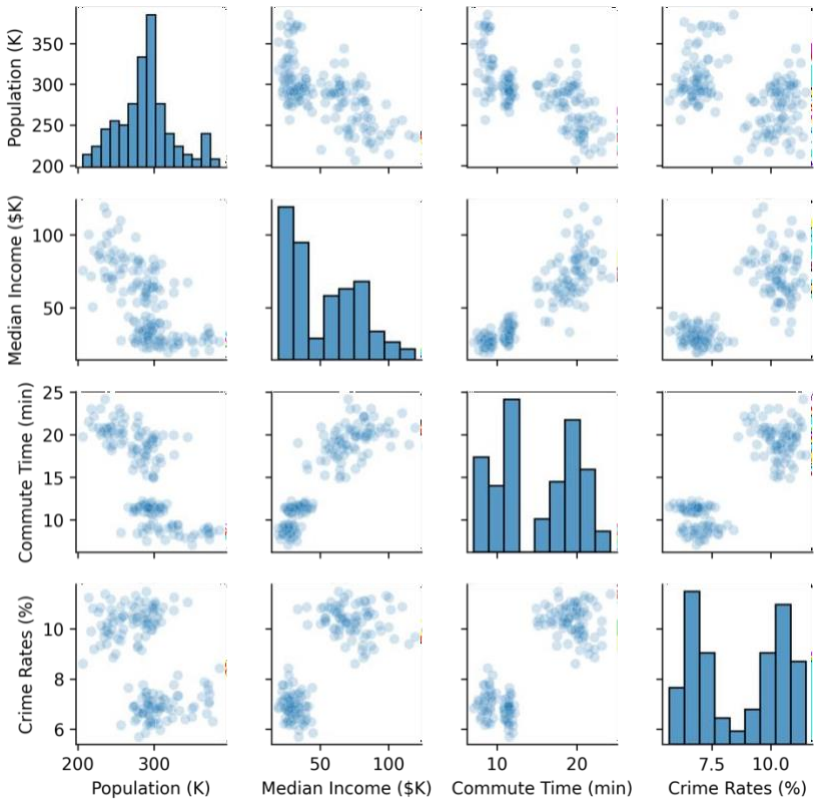



**Figure 4–1:** A matrix scatter plot of all features to understand the bivariate space and the relationships between each socio-economic factor. Where population is negatively correlated with the median income and commute time. Meanwhile, for crime rates we can see clear segmentations and groupings within its relationship to all other factors indicating a scenario of mixing populations; if left unattended can lead to the statistical Simpson's paradox. However, there seems to be a strong linear association between median income and commute time to work for the 158 communities considered.

Next, the dataset is normalized on [0.01, 1] and statistical outliers are checked for if any using the interquartile range method. Demonstrating the workflow, we obtain the sorted dissimilarity matrix of normalized predictors as the MDS input, which is used to obtain the dataset's projections in the low dimensional space shown in Figures 4–2 and 4– 3 respectively.

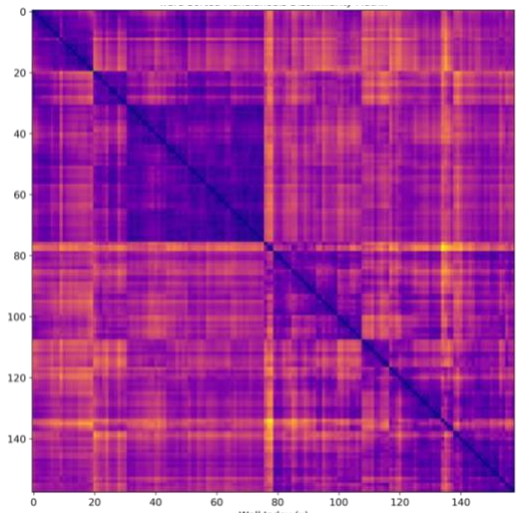


**Figure 4–2:** Sorted dissimilarity matrix showing 4 natural groupings between communities considered within the dataset using agglomerative hierarchical clustering with Ward's linkage.

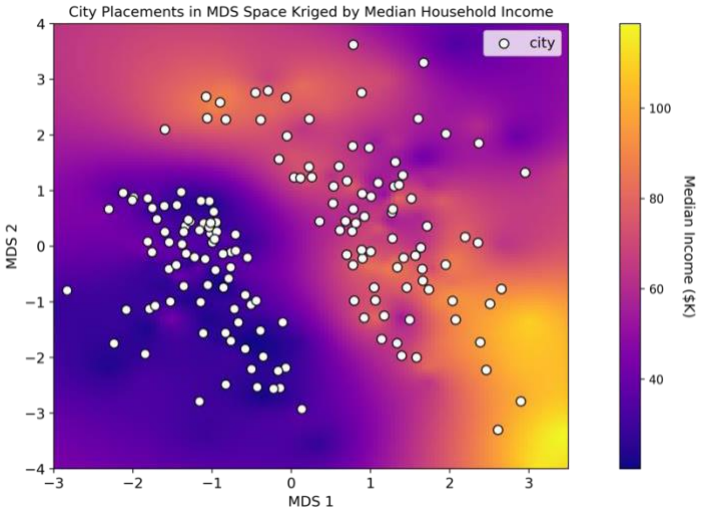



**Figure 4–3:** A kriged spatial map of median household income for different Texan communities over the equivalent area of interest in the subspace/ reduced dimension as indicated by the projections MDS 1 and MDS 2.

At a glance, Figure 4–3 shows two regions of interest, where the cold spots (dark blue) indicate communities that have an aggregated low median household income, and the hot spots (yellows & oranges) indicate sets of communities with relatively high household income. Ocularly, we see a clear divide between the communities that splits the data into 2 major groups. Although we can see some obvious trends and make evident inferences, we cannot necessarily determine or identify clusters from the data without the use of an efficient clustering algorithm.

Next, the DBSCAN clustering algorithm was implemented with its results shown in Figure 4–4 where 4 main community groups highlighted by the colors blue, magenta, green, and yellow are identified as clusters 1 through 4 respectively. Also, the proportion of the clusters to the entire dataset is 0.45, 0.30,0.025, and 0.055 for clusters 1 through 4 respectively. Meanwhile, the outlier label highlighted in black has a proportion of 0.17 indicating that 27 communities cannot be classified into the 4 main community archetypes found.

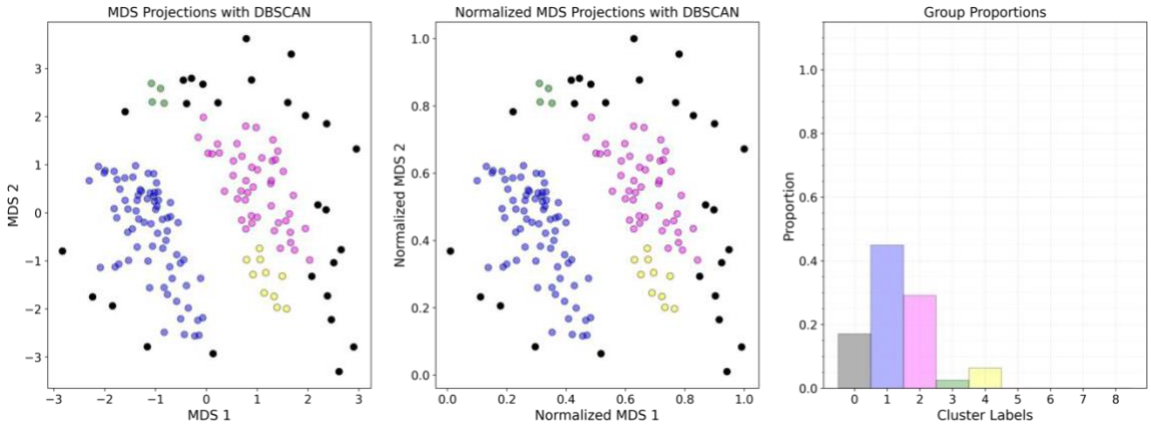
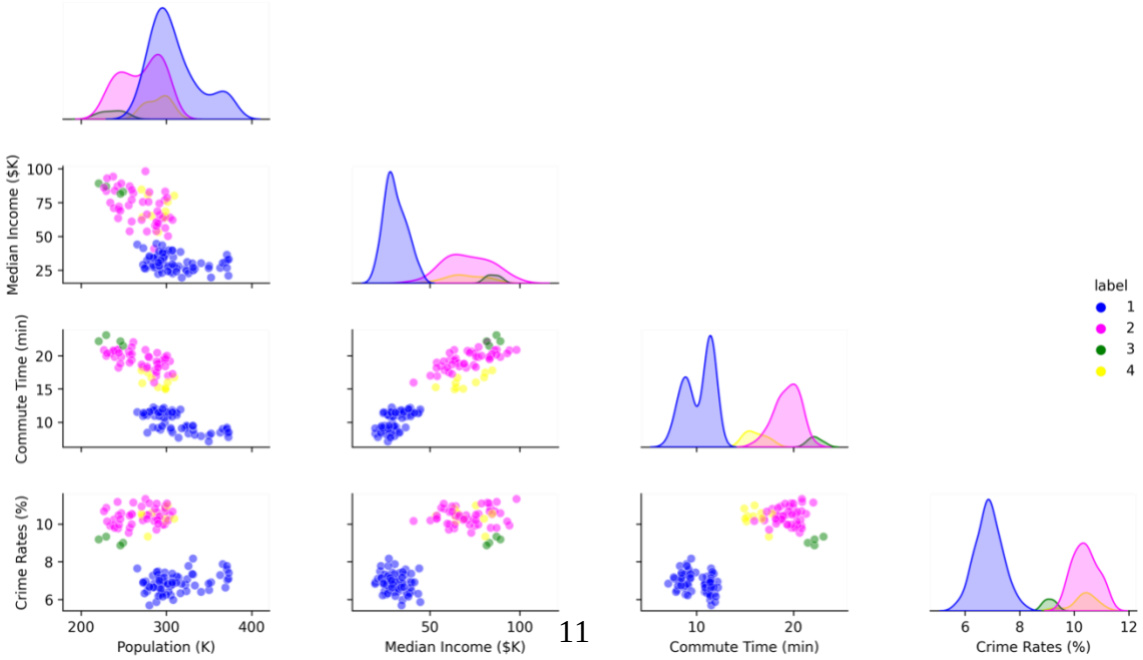



**Figure 4–4:** Left to right shows the clustered result of the communities in the MDS space, normalized MDS space, and the proportion of clustered groups found in the entire dataset respectively. The scatterplot of these communities shows 4 clusters as highlighted with the blue, magenta, green, and yellow colors representing cluster labels 1 through 4 respectively. Note: the cluster label highlighted in black with index 0 represents outlier cities as identified by DBSCAN is not statistically recognized as an outlier rather, the clustering method identifies these cities as such due to hyperparameter constraints – not having enough minimum points (min pts.) within the tuned epsilon (eps) as discussed in prior sections.

On finding the clusters, the matrix scatter plot is colored by each cluster to verify the initial hypothesis of mixing populations in the data (Figure 4–5). Upon close inspection, there is a clear distinction within the clusters found for all features alluding to the existence of subgroups in the data. An interesting find is that of crime rates (%) against commute time (minutes), where we see that erroneous inferences such as a negative linear association, can be made in the subgroup located on the top-right



as opposed to the actual positive association when clustered.

**Figure 4–5:** A matrix scatter plot colored by the 4 clusters found for all features to understand the bivariate space and the relationships between each socio-economic factor, where K represents thousands.

Figure 4–6 shows a prediction model to ascertain the testability of the workflow is generated using a KNN classifier having a test size of 0.30 and a training size of 0.70 for a dataset consisting of 158 samples. Figure 4–6 has a dual value add-on premise: i) the visualization of the test case with an accuracy of 0.975 after KNN and an averaged F1- score of 0.94, ii) the identification of boundary cities between two or more cluster groups using as a form of uncertainty quantification for the classification schemes found.

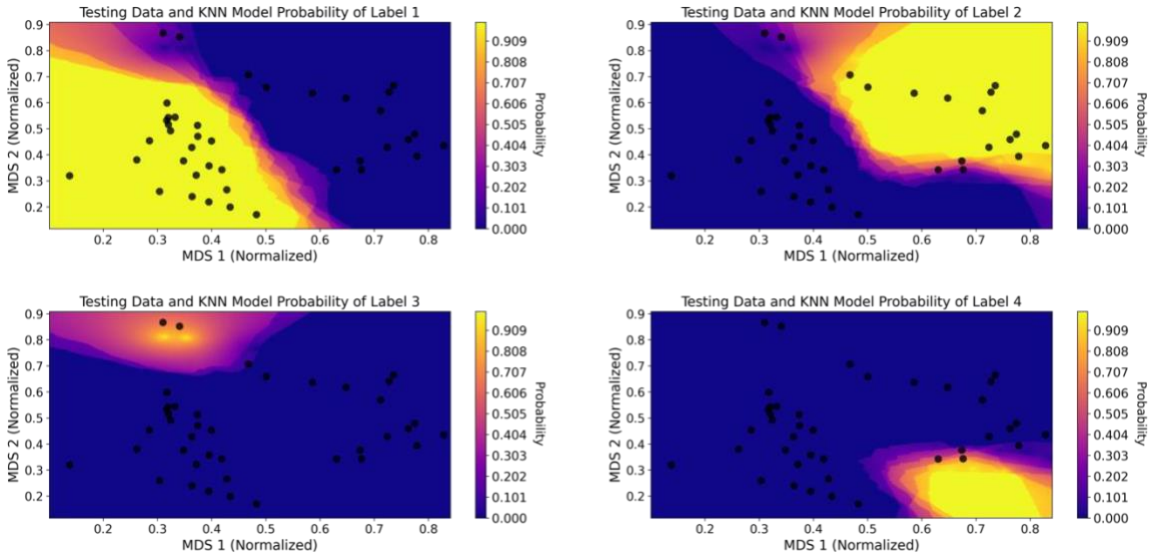



**Figure 4–6:** Visualization of the KNN predictive model for the test case showing with cluster identification to verify the labels found by DBSCAN. The background map represents the probability of a sample belonging to a particular cluster where the dark and bright colors represent regions ranging from low to high probability per the color bar. Also, these maps show the decision boundaries found by KNN for each of the identified cluster labels. Based on the subplots in the second column, we can see 3 data points on the boundary between clusters 2 and 4 having an approximate probability ranging between 0.20 and 0.35 of belonging to cluster 2.

After cluster identification and predictive model building for our Texan communities in the low dimensional MDS space, we find our analog communities to assist with comparative analysis when limited information is available about a particular community of interest. Figure 4–7 shows the entire sample size colored by the clusters found inclusive of the DBSCAN outlier group in the Euclidean space, which represents the high dimensional space with the multiple predictor features underlain by the kriged spatialmap of median household income on the left.

Next, the centroid of each cluster in the MDS space is computed yielding 4 analogs shown in Figure 4–7 on the right, which serves as a global representation of the entire dataset into 4 cities that can be used

for dissimilarity or similarity comparisons between individual communities based on the predictors. Lastly, the group consistency score (GCS) of select cities relative to the individual analogs as shown in Table 4–1, showed that communities with similar GCS scores across all clusters belong to the dubbed "non- statistical outlier" grouping based on the DBSCAN algorithm workings. Hence, we can infer that such results for the GCS determined for the specific test communities are considered inconclusive as its ranking and grouping may change over time or with more data.

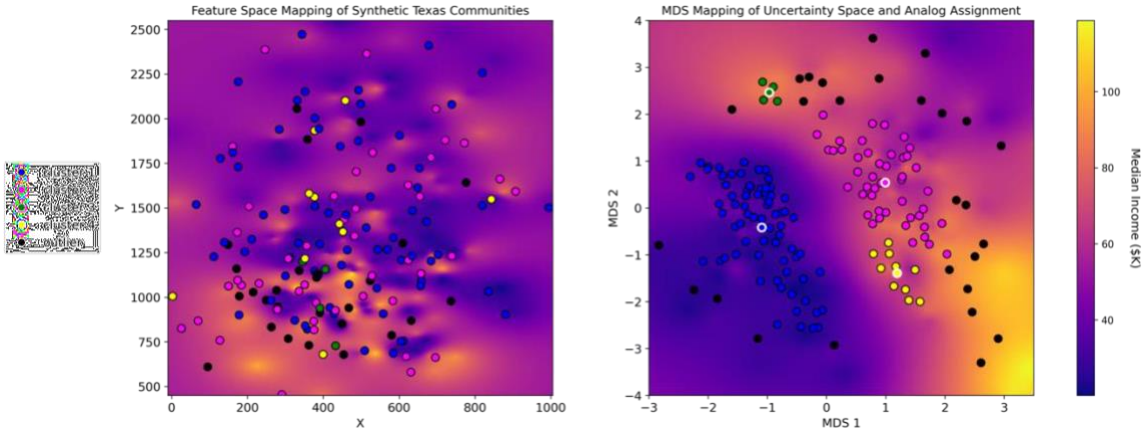



**Figure 4–7:** A similarity-based categorization of the communities and identification of 4 community clusters with samples colored by the clusters found in the MDS space with no evident rationale due to the curse of dimensionality on the left and its analog representations encircled in white on the right. Each of these clusters consists of communities with similar socio-economic attributes that help understand the feature of interest– median household income, in rural Texas. Note that communities classified as outliers are not statistical outliers but are samples that did not fit the major clusters based on insufficient density in the DBSCAN algorithm.

**Table 4–1:** *Group Consistency Score (GCS) for the 4 Select Communities at Random.*

| Index 6 | Index 54 | Index 102 | Index 140 |
|---|---|---|---|
| 0.16 | 0.45 | 0.28 | 0.29 |
| 0.27 | 0.42 | 0.08 | 0.56 |
| 0.35 | 0.57 | 0.20 | 0.63 |
| 0.47 | 0.76 | 0.42 | 0.66 |
| 0.27 | 0.35 | 0.45 | 0.02 |

## 5. Conclusion

Overall, one of the key benefits of our workflow is its ability to identify items or resources, in this case, Texan communities, and their socio-economic attributes as a measure of resilience as clusters, resulting in optimal decision-making during funds allocation. Other advantages include: i) identifying communities/ cities with similar socio-economic properties, ii) Mappings for uncertainty exploration and introduction of a probabilistic classification scheme, iii) decision making when faced with limited information and high uncertainty, and iv) the ability to provide sufficient flexibility to be adapted to different interdisciplinary needs while catering

to the statistics and spatial settings of the underlying system.

## References

Bhushan, V., & Hopkinson, S. C. (2002, October). A novel approach to identify reservoir analogues. In European Petroleum Conference. OnePetro.

Bygdevoll, J. 2007. How to Find Field Candidates for Enhanced Recovery by Water Additives on the NCS. Enhanced Recovery by Water Additives FORCE Seminar accessed April 28, 2021. https://www. yumpu. com/ en/ document/ view/ 38799117/ how- to- find- field-candidates- for- enhanced- recovery- by water- force.

Cox, D. R. (1955, July). The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 51, No. 3, pp. 433-441). Cambridge University Press.

Dromgoole, P. and Speers, R. 1997. Geoscore: A Method for Quantifying Uncertainty in Field Reserve Estimates.
Pet Geosci 3: 1–12. doi: https:// doi. org/10. 1144/ petgeo. 3. 1. 1.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

Gomes, J., Narayanan, R., Parra, H. et al. 2018. Benchmarking Recovery Factors from Carbonate Reservoirs: Key Challenges and Main Findings from Middle Eastern Fields. Paper presented at the Abu Dhabi International Petroleum Exhibition and Conference, Abu Dhabi, UAE, 7–10 November. SPE-193094-MS. doi: https:// doi. org/ 10. 2118/ 193094-MS.

Hodgin, J. E. and Harrell, D. R. 2006. The Selection, Application and Misapplication of Reservoir Analogs in the Estimation of Petroleum Reserves. Paper presented at the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, USA, 24–27 September. SPE-102505-MS. https:// doi. org/ doi: 10. 2118/ 102505- MS

Köppen, M. (2000, September). The curse of dimensionality. In 5th Online World Conference on Soft Computing in Industrial Applications (WSC5) (Vol. 1, pp. 4-8).

Kruskal, J.B.. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29:1–27, 1964.

Masoudi, R., Jalan, S., and Sinha, A. K. 2020. Application of a Novel Hybrid Workflow with Data Analytics and Analog Assessment for Recovery Factor Benchmarking and Improvement Plan in Malaysian Oilfields. Paper presented at the SPE Asia Pacific Oil & Gas Conference and Exhibition, Virtual, 17–19 November. SPE- 202459- MS. https:// doi. org/ doi: 10. 2118/ 202459- MS.

Olukoga, T. A., & Feng, Y. (2021). Determination of miscible CO2 flooding analogue projects with machine learning. Journal of Petroleum Science and Engineering, 109826.

Rodriguez, H. M., Escobar, E., Embid, S., Morillas, N. R., Hegazy, M., & Lake, L. W. (2014). New approach to identify analogous reservoirs. SPE Economics & Management, 6(04), 173-184.

Scheidt, C., & Caers, J. (2009). Representing spatial uncertainty using distances and kernels. Mathematical Geosciences, 41(4), 397-419.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 1-21.

Sidle, R. E. E. and Lee, W. J. J. 2010. An Update on the Use of Reservoir Analogs for the Estimation of Oil and Gas Reserves. SPE Econ & Mgmt 2 (2): 80–85. SPE-129688- PA. https:// doi. org/ 10. 2118/ 129688- PA.

Smalley, P. C., Ross, A. W., Brown, C. et al. 2009. Reservoir Technical Limits: A Framework for Maximizing Recovery from Oil Fields. SPE Res Eval & Eng 12 (4): 610–629. SPE-109555- PA. https:// doi. org/ 10. 2118/ 109555- PA.

Sun, S., & Pollitt, D. A. (2021). An Empirical Analog Benchmarking Workflow to Improve Hydrocarbon Recovery. SPE Reservoir Evaluation & Engineering, 1-18.

Tan, X., Tahmasebi, P., & Caers, J. (2014). Comparing training-image based algorithms using an analysis of distance. Mathematical Geosciences, 46(2), 149-169.