Enhancing Power Flow Studies Through Representative Scenario Selection

Jordan Cook, Maximo Briones, Farnaz Safdarian, Thomas Overbye
Texas A&M University
Department of Electrical and Computer Engineering
College Station, TX
{jordancook, max25, fsafdarian, overbye}@tamu.edu

***Abstract*—This paper presents an approach to scenario selection with the goal of improving the accuracy of power flow simulations, particularly with vast datasets involving load and weather variables. With large power systems and large amounts of available data, it is computationally expensive to choose important scenarios with a higher impact on the operation, considering load and weather for renewable generation output. Using the K-Means method for clustering, representative points are strategically chosen to simulate various solar, wind, and load conditions. The two selected representative points include an average and an outlier. Choosing these two points allows for baseline data analysis as well as anomalies, which can cause stress in the grid. The method is then demonstrated in this paper to show its functionality and how it captures the diversity of a dataset. The resulting clusters help finding interesting scenarios by addressing the variability that is inherent in power systems. This leads to improving grid reliability by preparing for a range of scenarios.**

***Index Terms*—Power flow studies, scenario selection, K-means clustering, Weather data, load data, FERC-714, electric grid resilience**

I. Introduction

When conducting planning studies on the electrical grid, both load and renewable generation vary throughout the year. Typically, the grid witnesses its peak demand for load during the summer, with winter ranking second in terms of high demand. On the other hand, load is typically the lowest in the spring. The traditional approach to scenario usage involves utilizing the peak load scenarios in summer and winter, along with the low load scenarios in spring. However, this paper departs from the traditional approach by introducing a novel method for identifying scenarios to be examined in planning studies, moving away from the conventional focus on specific operating points.

it is essential to align the level of realism with the specific objectives of the planning study when running power flow simulations. In the past, weather has only been modeled implicitly. This includes modeled load values, real and reactive power output, transmission line limits, transformer limits, and more.

Weather has long been affecting the grid and although the implicit methods were adequate in the past, weather data now needs to be modeled explicitly. The two primary reasons for this shift is a large increase in amount of renewable generators (specifically wind and solar) with the continued expansion of their capacity and the occurrence of extreme weather events with the need to plan for worst-case scenarios [1].

Incorporating weather data into power flow simulations is just one characteristic of the evolving landscape in grid modeling. Equally important is the inclusion of realistic load values and trends, reflecting the dynamic nature of power demand in various regions. The demand for electrical power exhibits fluctuations influenced by a multitude of factors. One of these is the significant impact that weather has on the load in an electrical grid [2]. For example, during the summer in southern areas of the United States, there is a surge in electricity due to air conditioning usage. On the other hand during the same season in more northern areas, there is less of a surge because it stays relatively cooler. Another factor is the economic or commercial operations in an area. Industrial hubs exhibit distinct load profiles compared to residential areas. Understanding these nuances is vital to understanding the infrastructure [3].

Another crucial consideration is the utilization of actual data in these simulations. For the most authentic and reliable outcomes, it is imperative to employ genuine, real-world data. The research in [4] also uses publicly available data to create synthetic time series of load at buses in a synthetic system. Additionally, [5] also develops scenarios by determining generation and load characteristics. Similar to these papers, this paper also uses publicly available data to create scenarios; the difference here is that a clustering method is used to reduce the amount of scenarios to be tested.

The research in [6] highlights the importance of clustering in reducing the amount of scenarios to aid with computational complexity. The research in [7] and [8] also mention the benefits in the context energy and power, with the both considering unit commitment and the second highlighting the wind generation scenarios. [9] compares several methods to reduce the amount of tested scenarios and does a case study on a 24-bus case.

The work in this paper presents a new method of three-dimensional clustering to study wind generation, solar generation, and load demand and choose the scenarios for testing. This is different from other clustering methods, as it will choose the central scenario as well as the outlier. There is a great need to study outliers because those scenarios are when the grid is under stress (the importance of outlying data in [10]). The scenario selection method will be demonstrated using Texas data for its renewable generation capacity and its unique grid situation. This method is aimed at enhancing the resiliency of the power grid by offering scenarios for testing in order to plan and prepare for all possible events.

Accepted to TPEC 2024, Texas, February 2024.

II. Sources of Data

A. Weather Data

Rather than relying on the output from renewable generators, as is commonly practiced in the literature, incorporating raw weather data directly into operational problems, such as Optimal Power Flow (OPF) as mentioned in [11], can enhance the precision of renewable generation estimates, track sudden weather changes, and provide more detailed insights. The proposed approach for directly integrating weather data into OPF employs conventional methods like Newton-Raphson, without significantly increasing the computational complexity.

This paper integrates detailed weather measurements, such as temperature, wind velocity and direction, humidity through dew point, solar radiation, and cloud cover, into the assessment of wind and solar generation capacities on an hourly basis, following the strategy mentioned in [11]. It harnesses weather data from 1973 to 2022 obtained from various weather stations across the continental United States, mapping this data to the generators based on geographic proximity. In instances of missing weather data, the paper elaborates on the use of Delaunay Triangulation for interpolating the adoption of data from the nearest station with complete records.

Historical weather data, is collected at an hourly granularity from various sources worldwide. [12], [13] The International Civil Aviation Organization (ICAO) and the World Meteorological Organization (WMO) serve as primary providers of meteorological data, which encompasses thousands of weather stations across the globe. Additionally, electric utilities may supplement this dataset. For instance, real-time weather data from weather stations, identified via ICAO codes, is accessible at [14].

Public datasets from the U.S. Energy Information Administration (EIA-860) [15] offer extensive details on the United States' power generators, including fuel type, capacity, precise locations, and specific attributes for wind and solar generators. These datasets enable the categorization of wind turbines and solar cells with similar features into power plants grouped by location.

For assessing weather impacts on renewable generators, six distinct models were developed. The first model estimates wind power plant output using variables such as local wind speed and turbine power curves, referencing [16], [17], and [18]. Models two through four, drawing on [19], are variations of the first, tailored to different wind turbine types. The fifth model projects solar PV output using local solar data and the configurations mentioned in [16], while the sixth model, informed by [20], predicts thermal generator output changes due to temperature variations. The output values of renewable resources based on this method are validated in [21].

For this study, the raw weather data was input into PowerWorld software [22] and thus the power output for each hour was calculated. From here, it was organized into solar and wind generation output for each state. This allows for studying the renewable generation at any given hour going back to 1973 separated by state.

B. Load Data

The Federal Energy Regulatory Commission (FERC) collects comprehensive data on the wholesale electric market. One of these datasets is FERC Form No. 714 [23], which is required for every electric utility with a planning area having an annual peak demand for power 200 MW or more. The form contains a section which reports the each planning area's hourly demand for each day of the year.

There are several ways that this data can be studied. The first way is to study each value directly with the MW value (this is ideal when only studying a certain area). Another way to is take all the data and normalize the value based on the maximum. This will give each datapoint a value between 0 and 1, with 1 being the peak load demand that area experienced in the year and each other value in the dataset is a ratio of the maximum. This method is more useful when doing studies across different areas. For example, if a researcher wants to compare the load demand in an area with high population (such as the SPP or MISO footprint) and want to compare the patterns in load usage to a smaller area (such as FRCC or WAPA), one can do this because all the values are between 0 and 1 instead of the actual MW values.

The following images show an example why this is helpful. For figure 1, one can see from the legend in figure 3 that SERC in grey, FRCC in orange, and the Northwest in purple are experiencing a high load demand, while CAISO in blue is not near its peak. Figure 2 has fairly average load demands in most areas, while FRCC in orange is experiencing a much higher demand comparatively. These comparisons would be difficult to make without normalizing the data between 0 and 1, as the areas all have a difference magnitude of load demand without their region.

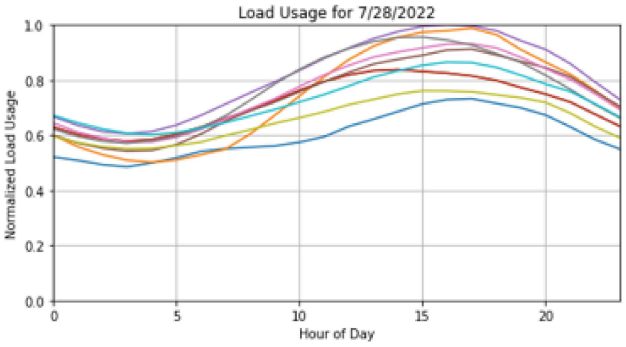

Fig. 1: Load trends for July 28, 2022 in different US areas

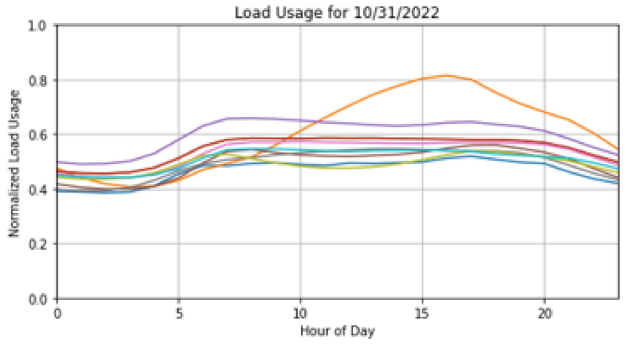

Fig. 2: Load trends for October 31, 2022 in different US areas

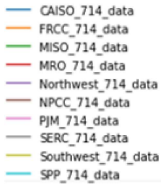

Fig. 3: Legend

III. The Problem

A common challenge faced with such vast datasets is that there is an abundance of information, but difficulty in finding meaningful representation of what is presented. For instance, studying a single year on an hourly basis yields 8,760 data points. When encompassing multiple factors for study, the number of potential scenarios grows exponentially. To illustrate, the combination of a year's worth studying hourly of two dimensional data leads to 80 million permutations for analysis. This complexity is compounded by the inclusion of decades-long datasets, dating back to 1972 for weather data and the early 2000s for load data.

This sheer magnitude of data points presents a challenge, due to computational limitations and resource constraints (time, amount of people working, available computers, etc.). The need for a discerning approach in data point selection arises from the practical need to optimize the testing process and work efficiently, ensuring a focused and meaningful analysis within the constraints of available resources.

In the past, the load values at the summer peak are often used when running these simulations, as well as scenarios in the winter or the spring for lows. Although it is important to plan for demand that would cause the grid stress, it is necessary to consider weather, as the dependency on renewable generation grows. This approach also couples wind and solar weather data with the load to get accurate, realistic scenarios to run as operating points.

IV. The Proposed Solution

K-Means Clustering, as explained in [24], is a popular unsupervised machine learning algorithm that aims to group data points together based on their characteristics. The solution proposed here uses this K-Means clustering method in three dimensions to aid in data point selection. This method uses three characteristics:

1) Calculated Output of Solar Generation
2) Calculated Output of Wind Generation
3) Total Load

For every n-clusters, there will be $2n$-scenarios for testing. This is because two points will be chosen from each cluster - the point closest to the centroid and the point furthest away from the centroid ('closeness' refers to the Euclidian distance in three dimensions). This allows for both average data to be studied as well as the outlying data.

The inclusion of both average data points and outlying data points is crucial for obtaining a comprehensive understanding of the underlying patterns and dynamics within a dataset. Analyzing average points allows for grasping central tendencies and typical behaviors exhibited by the features studied in the power grid, therefore providing a baseline understanding of the general trends. On the other hand, the examination of outlying points is equally essential, as these instances often carry valuable information about anomalous conditions. The study of outliers aids in identifying potential scenarios when the grid faces periods of stress. Typically if a grid can withstand the troublesome times, it can withstand baseline operations. Thus, the dual exploration of averages and outliers enhances the robustness of this analysis, offering a more complete perspective.

When choosing the amount of scenarios that should be studied, it is important to also consider the data properties. As explained in [25], the elbow method is a way to determine the minimum number of clusters a dataset should have. When calculating the variance that is preserved for an increasing number of clusters, there is a point in the curve, known as an 'elbow' or 'knee' that is a leveling off-point. The integer that corresponds to this change is often considered the minimum number of clusters that is ideal for a dataset. It is important to note that the 'elbow' point does not exclude the need to have more clusters, as the amount that is chosen for each study must reflect the necessary level of depth needed to understand a situation.

Since this method gives a data point that is only a singular time snapshot, it would also be beneficial to conduct studies based on the entire day for the selected time point. This is particularly helpful when trying to study how the grid morphs or changes based on past and future conditions. This is only something to keep in mind based on the studies that are being performed.

V. Case Study

To effectively demonstrate the method proposed in this paper, a case study will be showcased utilizing Texas data. The choice of Texas is particularly apt, given its distinctive characteristics as both a self-contained grid with large renewable penetration and a singular state. This allows us to have consistent data from one entity, which significantly reduces the amount of discrepancies in the time-series data used.

For this case study, data from 2021 was used since that is the latest publicly available load data provided by FERC. Although for each study, this can be tailored as necessary, even concatenating data sets that span multiple years or decades. A single year is easier to demonstrate and visualize for the purposes of this paper.

This raw load and generation data can be seen in Fig 4. Something to note is that solar generation is relatively low compared the wind generation and total load, so pre-processing has to be done on the raw data so that the algorithm takes the solar generation input into account (such an normalizing the data so that all variables are given equal consideration in the clustering process). The different scales of solar, wind, and load can be seen in Figure 7, as each box and whisker plot shows the MW values from the dataset.

From here, the elbow method is used. The following result can be seen in Figure 5, which leads to a selection of 4 clusters to capture the data, as that is the integer that has the sharpest angle change.

As seen in Figure 6, the data is split into four base clusters. The characteristics of these clusters can be generalized in Table I.

It is important to mention that there is only one cluster for the case of high solar generation independent of total load and wind generation due to the fact that there is a relatively low density of data points with high solar generation. This

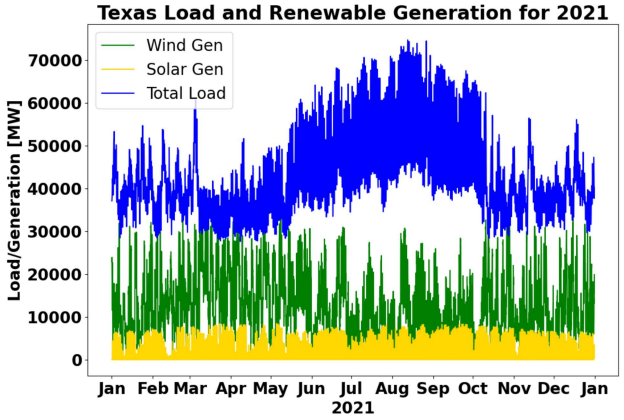

Fig. 4: Texas Load and Generation 2021

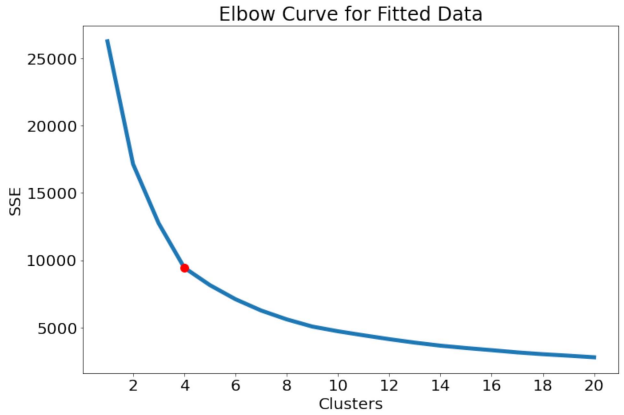

Fig. 5: Elbow Plot for Texas 2021 Data

is because solar has many points around zero because of the nature of solar generation at night. Wind varies around much, going from 0 MW to upwards of around 30,000 MW. Load never dips below 25,000 MW for any time points, so this also is represented in the clusters.

Finally looking at Figure 8, one can see X's that mark the data that is closest to the centroids of the clusters. On the other hand, the diamonds that mark the points that are furthest from their respective centroid. These furthest points are the extremes of typical operating conditions that could potentially affect the stability of the grid. These selected points reduce the amount of data points that we we have to simulate, while still having a diverse range of scenarios.

Although the centroid in the green cluster is close to the pink cluster, it is still a scenario that merits testing, especially since no other points in its proximity are being tested. Both the pink and teal outlier point will merit interesting scenarios that would not normally be tested when studying typical grid operations.

Figure 9 shows a bar plot of the different scenarios, with the centroid on top being directly compared to the outlier in each cluster. This allows one to visualize and easily understand what is being changed in each of the proposed scenarios.

VI. Conclusion

This paper showed an initial approach for scenario selection using K-Means clustering. It used both average points

TABLE I: Cluster Characteristics

Color	Solar	Wind	Load
Green	Low	Low	Low
Gray	Low	High	Low
Pink	Low	High	High
Teal	High	High	Varies

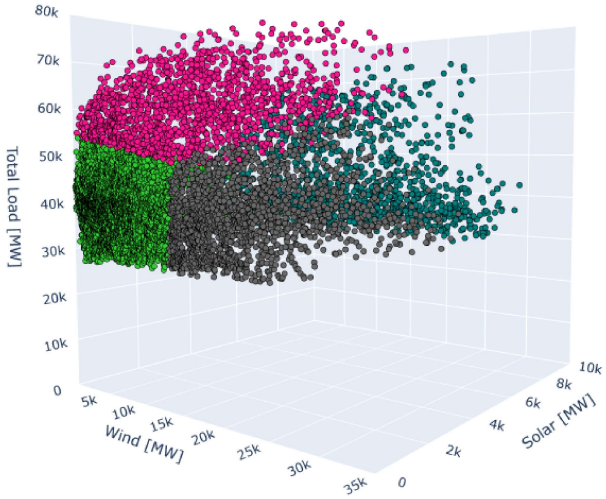

Fig. 6: 3D Clustering of Data

of operation that is often seen as well as the outliers that could potentially cause some stress in the grid. This method was demonstrated and visualized using data from 2021 in the state of Texas.

The results help finding interesting scenarios based on using clusters discovered in the raw data. Future work includes adding more dimensions into this study, such as time itself, region, or more weather factors. Adding time as a dimension for clustering could potentially help in finding scenarios for different seasons, adding regions could help in identifying areas of weakness in a grid, and adding more weather measurements would make the model more accurate. Validation could also be done by studying these selected points in PowerWorld and seeing the effects these scenarios have on the grid.

Acknowledgment

This work was partially supported through funding provided by the Power Systems Engineering Research Center (PSERC) through project S-99 and partially by Advanced Research Projects Agency–Energy (ARPA-E) for grid optimization projects.

References

[1] I. Staffell and S. Pfenninger, "The increasing impact of weather on electricity supply and demand," *Energy*, vol. 145, pp. 65–78, 2018.

[2] T. Hong, W.-K. Chang, and H.-W. Lin, "A fresh look at weather impact on peak electricity demand and energy use of buildings using 30-year actual weather data," *Applied energy*, vol. 111, pp. 333–350, 2013.

[3] E. Veldman, M. Gibescu, H. J. Slootweg, and W. L. Kling, "Scenario-based modelling of future residential electricity demands and assessing their impact on distribution grids," *Energy policy*, vol. 56, pp. 233–247, 2013.

[4] H. Li, J. H. Yeo, A. L. Bornsheuer, and T. J. Overbye, "The creation and validation of load time series for synthetic electric power systems," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 961–969, 2021.

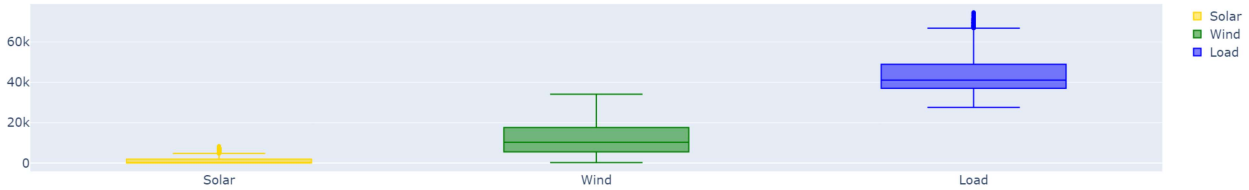

Fig. 7: Box & Whisker Plot for 3 variables

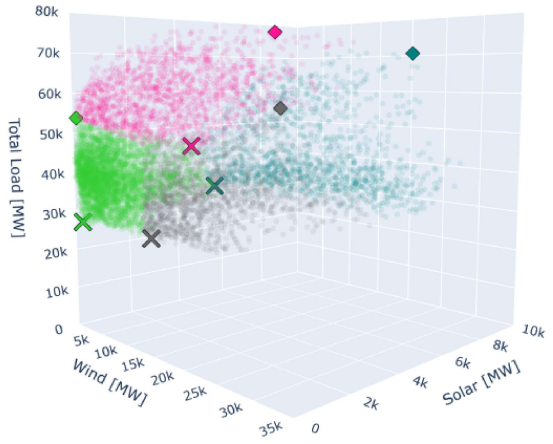

Fig. 8: Closest and Furthest Data Points

[5] H. Li, J. H. Yeo, J. L. Wert, and T. J. Overbye, "Steady-state scenario development for synthetic transmission systems," in *2020 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 2020, pp. 1–6.

[6] J. Hu and H. Li, "A new clustering approach for scenario reduction in multi-stochastic variable programming," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3813–3825, 2019.

[7] H. Keko and V. Miranda, "Impact of clustering-based scenario reduction on the perception of risk in unit commitment problem," in *2015 18th International Conference on Intelligent System Application to Power Systems (ISAP)*. IEEE, 2015, pp. 1–6.

[8] E. Du, N. Zhang, C. Kang, J. Bai, L. Cheng, and Y. Ding, "Impact of wind power scenario reduction techniques on stochastic unit commitment," in *2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO)*. IEEE, 2016, pp. 202–210.

[9] Y. Dvorkin, Y. Wang, H. Pandzic, and D. Kirschen, "Comparison of scenario reduction techniques for the stochastic unit commitment," in *2014 IEEE PES General Meeting— Conference & Exposition*. IEEE, 2014, pp. 1–5.

[10] J. W. Osborne and A. Overbay, "The power of outliers (and why researchers should always check for them)," *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, p. 6, 2004.

[11] T. J. Overbye, F. Safdarian, W. Trinh, Z. Mao, J. Snodgrass, and J. H. Yeo, "An approach for the direct inclusion of weather information in the power flow," *Proc. 56th Hawaii International Conference on System Sciences (HICSS)*, 2023.

[12] "ERA5 hourly data on single levels from 1940 to present". [Online]. Available: https://cds.climate.copernicus.eu/cdsapp!/dataset/reanalysis-era5-single-levels?tab=form

[13] "Weather Station Identifiers". [Online]. Available: http://www.weathergraphics.com/identifiers/

[14] "Current Weather and Wind Station Data". [Online]. Available: https://aviationweather.gov/adds/dataserver_current/current/metars.cache.csv

[15] (2021) "U.S. Energy Information Administration (EIA)". [Online]. Available: https://www.eia.gov/electricity/data/eia860/

[16] G. M. Masters, *Renewable and efficient electric power systems*. John Wiley & Sons, 2013.

[17] V. Sohoni, S. Gupta, R. Nema *et al.*, "A critical review on wind turbine power curve modelling techniques and their applications in wind based energy systems," *Journal of Energy*, vol. 2016, 2016.

[18] P. Giorsetto and K. F. Utsurogi, "Development of a new procedure for reliability modeling of wind turbine generators," *IEEE Transactions on Power Apparatus and Systems*, no. 1, pp. 134–143, 1983.

[19] C. Draxl, A. Clifton, B.-M. Hodge, and J. McCaa, "The wind integration national dataset (wind) toolkit," *Applied Energy*, vol. 151, pp. 355–366, 2015.

[20] A. De Sa and S. Al Zubaidy, "Gas turbine performance at varying ambient temperature," *Applied Thermal Engineering*, vol. 31, no. 14-15, pp. 2735–2739, 2011.

[21] J. L. Wert, T. Chen, F. Safdarian, J. Snodgrass, and T. J. Overbye, "Calculation and Validation of Weather-Informed Renewable Generator Capacities in the Identification of Renewable Resource Droughts," in *IEEE PowerTech 2023*, 2023.

[22] T. J. Overbye, P. W. Sauer, C. M. Marzinzik, and G. Gross, "A user-friendly simulation program for teaching power system operations," *IEEE Transactions on Power Systems*, vol. 10, no. 4, pp. 1725–1733, 1995.

[23] F. E. R. Commission, "Form no. 714 - annual electric balancing authority area and planning area report," December 2022.

[24] J. Ortega, N. Almanza-Ortega, A. Vega-Villalobos, R. Pazos-Rangel, J. C. Zavala-Diaz, and A. Martínez-Rebollar, *The K-Means Algorithm Evolution*, 04 2019.

[25] H. Humaira and R. Rasyidah, "Determining the appropiate cluster number using elbow method for k-means algorithm," in *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*, 2020.

Fig. 9: Centroid vs. Outlier