Automated Selection of Number of Clusters for Determining Proliferation Resistance Measures

Daniya Zamalieva[a], Zachary Jankovsky[b], Alper Yilmaz[a], Tunc Aldemir[b], and Richard Denning[b]

[a] Photogrammetric Computer Vision Laboratory, The Ohio State University, Columbus, OH, USA
[b] Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH, USA

Abstract: Analyzing possible proliferation scenarios can provide insights on the most vulnerable stages of a nuclear system. With large number of scenarios, their manual examination becomes infeasible. One possibility for reducing the complexity of the data and discovering possible trends is via automated grouping of scenarios. The *k*-means clustering algorithm is widely used to group large amounts of data. This algorithm is very efficient, however, it requires the number of clusters to be known a priori. In this paper, we aim to overcome this issue by investigating several goodness-of-fit measures. Namely, using a set of proliferation scenarios modeled by PRCALC, we implement and compare the Bayesian Information Criterion, Akaike Information Criterion, Cluster Cohesion Coefficient and Anderson-Darling Normality Test to estimate the optimal number of clusters *k* for the *k*-means clustering algorithm. Experiments show that the examined measures can provide insights on the structure of the data.

Keywords: Scenario grouping, model estimation, PRCALC, non-proliferation

1. INTRODUCTION

The identification of the stages of a nuclear fuel system that are the most vulnerable to material diversion is an important element of assuring proliferation resistance. Software developed at Brookhaven National Laboratory called PRCALC [1] is capable of modeling the proliferation process as a Markov chain to estimate various proliferation resistance measures. Because comprehensive coverage of the PRCALC parameter space may lead to thousands of scenarios, manual analysis of the resulting data is infeasible. Clustering the output data and then analyzing the properties of the clusters can yield insight into the most vulnerable stages of a fuel cycle.

Recently, there has been effort at The Ohio State University towards automated analysis of PRCALC output [2] by grouping the produced scenarios using various clustering techniques. A clustering algorithm referred to as *k*-means is one of the most popular clustering methods, widely encountered in literature. This method is very efficient, and, therefore, it is preferred for very large datasets, such as the PRCALC output. A drawback of *k*-means algorithm is that the number of clusters *k* must be known a priori. The number of clusters is generally determined based on prior knowledge or practical experience, which may not always be available and places an unnecessary burden on a PRCALC user. Moreover, the true number of clusters is often not obvious, especially when the data dimensionality is high. An improperly chosen value of *k* may have an adverse effect on the resulting grouping. Introducing the option to select the number of clusters automatically would provide new insights on the underlying structure of the data and increase the applicability of *k*-means algorithm.

Several algorithms that determine the number of clusters *k* automatically have been proposed in the literature. A common approach is to evaluate clustering outcomes with different number of clusters based on a certain scoring function. In this paper, we implement and compare several scoring functions used to evaluate clustering outcomes, namely the Bayesian Information Criterion (BIC) [3], Akaike Information Criterion (AIC) [4], Cluster Cohesion Coefficient (CCC) [5] and Anderson-Darling Test (ADT) [6]. BIC and AIC evaluate a given model based on the likelihood of the data fitting this model penalized by the model complexity. They try to balance the goodness of fit with the

function of the number of parameters used to describe the model. The CCC is based on the inter- and intra-cluster similarity, where the desired grouping results in high similarity within the same cluster and low similarity between the members of different clusters. The ADT is a powerful normality test based on the empirical cumulative distribution function. It is used to decide whether the members of the same cluster are sampled from a Gaussian distribution, so that the cluster is further split if the test fails. In addition, we compare *k*-means results to that of mean-shift [7] and adaptive mean-shift [8], which are alternative clustering methods.

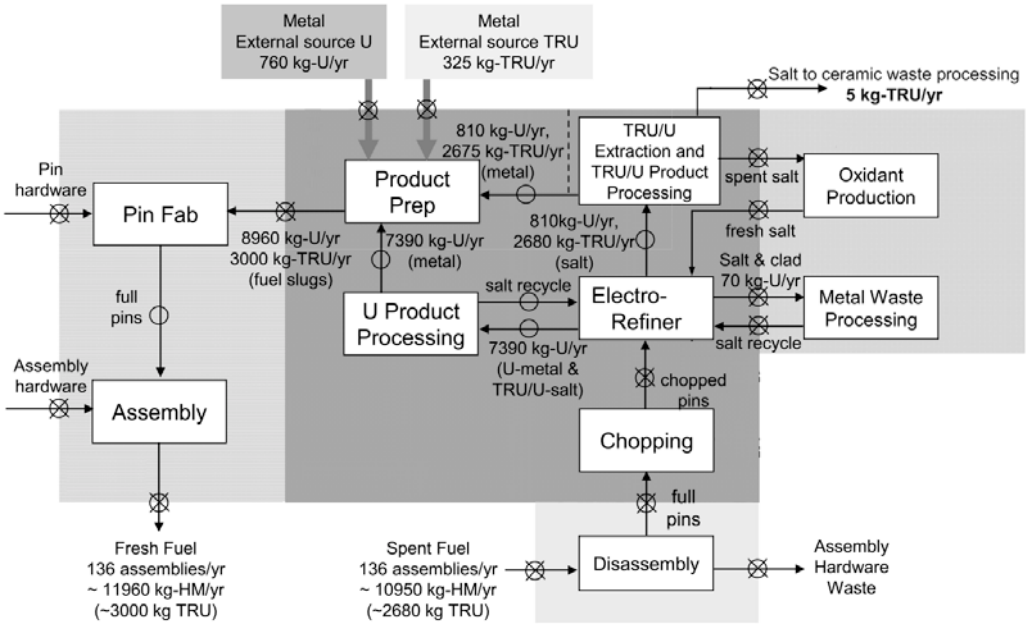
The rest of the paper is organized as follows. In the next section, we describe PRCALC and data generation process. Section 3 summarizes the methods for automated model selection. The results of the application of these methods to the generated data are presented in Section 4. Finally, we conclude in Section 5.

2. SCENARIO GENERATION

The Treaty on the Non-Proliferation of Nuclear Weapons, which went into force in 1970, is primarily enforced through safeguards implemented by the International Atomic Energy Agency (IAEA). There are various types of safeguards including inspections of records, surveillance of activity in sensitive areas, and material accountancy. A particular situation of interest in the field of proliferation resistance (PR) is a nation that allows safeguards at a nuclear site but covertly diverts material for illicit use.

PRCALC is a tool developed to address covert diversion scenarios. Its main focus is on a site comprising multiple Generation IV reactors for power generation as well as a fuel reprocessing plant. The hypothetical reactor is the Example Sodium-cooled Fast Reactor (ESFR), which takes in light water reactor (LWR) spent fuel (SF) as its main fuel source. LWRSF is composed mostly of uranium-238, uranium-235 fission products, and transuranic (TRU) elements. In the reprocessing plant seen schematically in Figure 1, the LWRSF is physically broken down and chemically separated. Some of the materials cannot be used in the ESFR, and are diverted to a waste stream. The useful fuel material is then re-formed into ESFR fuel assemblies for use in the reactors. The overall aim is to retrieve more energy from LWR fuel before it is considered waste.

Figure 1. PRCALC Reprocessing Plant Schematic (adapted from [9]).



In the scenario modelled by PRCALC, each step of the recycling process is a potential target for the would-be proliferator. Material diversion is represented by a Markov model, with 3 absorbing states: Success, Failure, and Detection. Success, from the point of view of the proliferator, is to obtain 1 significant quantity (SQ) equivalent of nuclear material, which the IAEA defines as, "the approximate amount of nuclear material for which the possibility of manufacturing a nuclear explosive device cannot be excluded." Failure represents a technical failure of some sort, either in obtaining material or in processing it to a usable form by the proliferator. Detection represents an alarm being raised and confirmed by the safeguards system. The probabilities of Detection (DP), Success (PS), and Failure (PF) are outputs of PRCALC, and sum to 1 for any given scenario. Another output is Proliferation Time (PT), which estimates the time that would be required, given the rates and locations of diversion, to obtain 1 SQ equivalent and process it into its form. The final output is Material Type (MT), which reflects the effort required to process the diverted material into usable form. For example, reactor-grade plutonium has a higher MT value than LWRSF.

PRCALC scenarios are unique combination of input parameters that are run to produce the outputs described in the previous section. For this work targets, diversion rates, and safeguard conditions were varied to create 65,520 scenarios. Seven targets were chosen (Table 1) out of the 23 potential targets in the PRCALC ESFR model. Four diversion rates were chosen: 0, 1σ, 2σ, and 4σ. σ refers to a fractional diversion rate of material from a target with an implicit increased probability of detection with higher values of σ. Four safeguard conditions were chosen, starting with the default safeguards in the PRCALC ESFR model. The remaining 3 conditions simulated the complete breakdown of Physical Inventory Verification (PIV), Surveillance & Monitoring, or Containment safeguards. These were included to show the effects of the proliferator surreptitiously compromising the integrity of IAEA safeguards.

Table 1. PRCALC Targets Chosen.

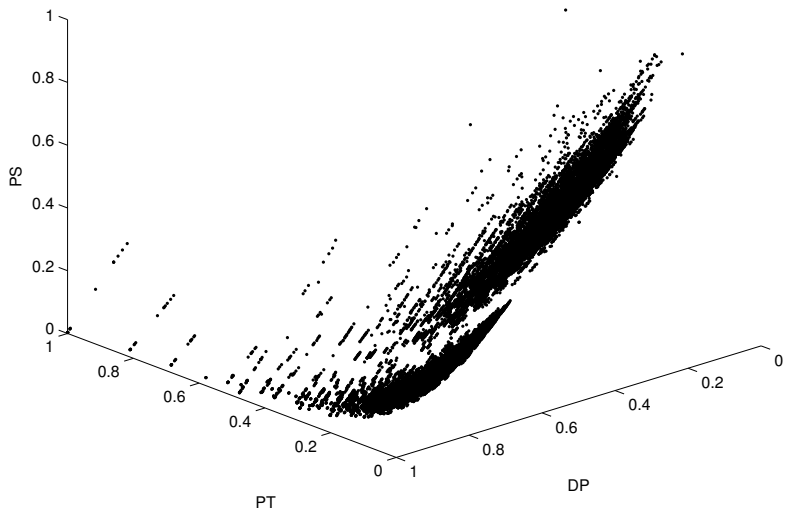
Target Number	Target Name
1	TRU Extraction
2	Electro-refiner
3	Pin Fabrication
4	Storage Basket: Fresh Fuel
5	ESFR SF Disassembly
6	Chopping
7	LWR SF Storage

Scenarios were created by taking every combination of the targets and diversion rates. Scenarios were then copied and modified for each safeguard condition. The number of scenarios was calculated as $N_{scen} = r^{N_{stage}} \times N_{SG}$, where N_{scen} is the total number of scenarios, N_{stage} is the number of target stages, r is the number of possible diversion rates, and N_{SG} is the number of safeguard conditions. This yields 65,536 scenarios. Four scenarios were created with a diversion rate of 0 for every target, and these scenarios were removed. Finally, 12 scenarios were removed before clustering due to having very high values of PT. These scenarios existed in 3 groups of 4 scenarios each, and were clearly outliers to be considered separately from the rest of the set. This left 65,520 scenarios to be clustered. The scenarios are visualized in Figure 2.

3. AUTOMATED MODEL SELECTION

In this section, we briefly describe the criteria employed in automating the selection of number of clusters k. We first introduce the notation used throughout the paper. The set of all observations is represented as a $n \times d$ matrix D, where n is the number of observations and d is the dimensionality of the feature space. The model M_k denotes the clustering result with k clusters $C_1, C_2, \ldots, C_k$. The number of observations in cluster C_i is referred to as n_i.

Figure 2. Visualization of generated scenarios. The dimensions correspond to PT, DP and PS.



3.1. Bayesian Information Criterion (BIC)

The BIC, also known as Shwarz criterion, can be computed as [3]

$$BIC(M_k) = \hat{l}_k(D) - \frac{p_k}{2}\log n, \tag{1}$$

where $\hat{l}_k(D)$ is the maximum log-likelihood of the data with respect to the model M_k, and $p_k = k(d+1)$ is the number of parameters in M_k. According to BIC, the best model is the one resulting in the highest log-likelihood and has the lowest number of parameters. Since *k*-means assumes spherical Gaussian for each cluster shape, the maximum likelihood estimate for the variance is

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_i (x_i - \mu_{(i)})^2 , \tag{2}$$

where $\mu_{(i)}$ is the centroid of the cluster to which x_i was assigned. The point probabilities can be computed using

$$\hat{p}(x_i) = \frac{n_i}{n}\cdot\frac{1}{\sqrt{2\pi}\hat{\sigma}^d}\exp\left(-\frac{1}{2\hat{\sigma}^2}\left\|x_i - \mu_{(i)}\right\|^2\right), \tag{3}$$

Then, the log-likelihood of the data with respect to the model M_k is calculated using the point probabilities as

$$\begin{aligned}\hat{l}_k(D) &= \log\prod_{i=1}^{n}\hat{p}(x_i) = \sum_{i=1}^{n}\left(\log\frac{1}{\sqrt{2\pi}\hat{\sigma}^d} - \frac{1}{2\hat{\sigma}^2}\left\|x_i - \mu_{(i)}\right\|^2 + \log\frac{n_{(i)}}{n}\right)\\ &= -\frac{n}{2}\log(2\pi) - \frac{nd}{2}\log(\hat{\sigma}^2) - \frac{n-k}{2} + \sum_{i=1}^{k} n_i \log n_i - n\log n .\end{aligned} \tag{4}$$

The BIC score is used to evaluate the clustering schemes with different *k,* and the model resulting in the highest BIC score is selected. In context of *k*-means algorithm, the BIC was applied by Pelleg *et al.* [10] to introduce a variant of the *k*-means algorithm referred to as *x*-means.

3.2. Akaike Information Criterion (AIC)

Similarly to the BIC, the AIC aims to balance the goodness of fit with the function of the number of parameters used to describe the model. The AIC has the form

$$AIC(M_k) = \hat{l}_k(D) - p_k\,, \tag{5}$$

where $\hat{l}_k(D)$ is defined in (4) and $p_k = k(d+1)$. Note that, with this formulation, the BIC penalizes the complexity of the model more heavily than the AIC. As with BIC, the higher AIC score indicated better model fit.

3.3. Cluster Cohesion Coefficient (CCC)

The CCC for a cluster C_i is defined as [5]

$$CCC(C_i) = \frac{1}{n_i}\sum_{j\in C_i}\frac{B_j - A_j}{\max(A_j, B_j)}, \tag{6}$$

where

$$A_j = \max_{p\in C_i}\left\|x_j - x_p\right\| \quad \text{and} \quad B_j = \min_{p\notin C_i}\left\|x_j - x_p\right\|, \tag{7}$$

so that A_j is the maximum distance between a given point and all other points within the same cluster, and A_j is the minimum distance between a given point and all points belonging to other clusters. The CCC for a model M_k is the computed as

$$CCC(M_k) = \frac{1}{k}\sum_{i=1}^{k} CCC(C_i)\,. \tag{8}$$

The low values of CCC indicate low coherence and, therefore, a poor model. The best model is selected as the one having the highest CCC score.

3.4. Anderson-Darling Test (ADT)

ADT leverages the fact that *k*-means algorithm assumes spherical distribution for each cluster. In other words, the data points in each cluster are assumed to be sampled from a multidimensional Gaussian distribution. The ADT is based on the Anderson-Darling statistic, which is "normality measure". Let $x_1, x_2, \ldots, x_n$ be the sorted data that have been standardized to have mean 0 and variance 1. Let $z_i = F(x_i)$, where F is the $N(0,1)$ cumulative distribution function. The Anderson-Darling statistic is then computed as

$$A^2(Z) = -\frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log(z_i) + \log(1 - z_{n+1-i})] - n. \tag{9}$$

It has been shown that when the mean μ and the standard deviation σ are estimated from the data, the statistic is further corrected as [11]

$$A_*^2(Z) = A^2(Z)(1 + 4/n - 25/n^2). \tag{10}$$

Hamerly *et al.* [6] proposed a variant of *k*-means algorithm referred to as G-means, which discovers an appropriate number of clusters k using ADT to decide whether to split an existing cluster into two or to stop if the cluster follows a Gaussian distribution. Let $M_q = \{C_1, C_2, \ldots, C_q\}$ be the initial set of cluster centers, with a small q. The G-means algorithm can be summarized as following:

1. Perform initial clustering with q, such that $M_q = kmeans(X, q)$.
2. For each $i = 1, 2, \ldots, q$
   2.1. Let $X_i = \{x_j \mid j = 1, 2, \ldots, n_i\}$ be the data that belongs to center C_i.
   2.2. Initialize two centers as $C_i \pm s\sqrt{2\lambda/\pi}$, where s is the main principal component of X_i with eigenvalue λ. Run *k*-means on X_i with these initial centers. Let the resulting centers be c_1, c_2.
   2.3. Compute a *d*-dimensional vector $v = c_1 - c_2$ connecting c_1 and c_2, which is the direction important for *k*-means clustering. Project X_i onto v as $x'_j = \langle x_i, v \rangle / \|v\|^2$ to obtain a 1-dimensional representation of X_i. Then, transform X'_i to have 0 mean and variance 1.
   2.4. Compute $Z = F(X'_i)$. If $A^2_*(Z)$ is within the range of non-critical values at a confidence level α, then keep the cluster C_i and continue to 2.1 with $i = i + 1$. Otherwise, keep c_1, c_2 in place of C_i and continue to 1 with $q = q + 1$.
3. Stop when no new clusters are added.

The significance level α is chosen according to Bonferroni adjustment to reduce the chance of incorrectly splitting the clusters over multiple tests [6].

4. EXPERIMENTS

To apply the BIC, AIC, CCC and ADT discussed in Section 3, we first perform *k*-means on the scenarios generated as described in Section 2. To reduce the effect of random initialization of *k*-means algorithm, we ran the algorithm 100 times for each *k* = 2, ..., 70. Figure 3 shows the plot of average BIC value for each *k* (shown in black). We further fit a polynomial curve (shown in gray) to the BIC values to eliminate the occasional peaks corresponding to local maxima. According to BIC, the best number of clusters for the data is *k* = 24. The same procedure is applied for AIC. Similarly to BIC, the highest AIC score results in *k* = 24 (Figure 4). Also observe that, in general, BIC and AIC curves appear very similar. Due to the large number of data points, the likelihood terms in BIC and AIC are significantly larger than the second terms (see Equations (1) and (5)), therefore, the difference in the second terms becomes negligible. Also note that both BIC and AIC tend to decrease gradually after *k* = 24, indicating that the clustering model is a poorer fit after that point. The clustering structure with 24 clusters is depicted in Figure 5. Points shown with identical symbol and color are the members of the same cluster.

Figure 6 shows the plot of the CCC values. Recall that the highest CCC value indicates the best model. While for the range of *k* = 2, …, 70 the highest CCC value is obtained for *k* = 2, we observe the increasing trend after around *k* = 15, which indicates that the best *k* according to CCC score might be greater than 70. To confirm this assumption, we compute the CCC for *k* = 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 which results in CCC = -0.8147, -0.7802, -0.7604, -0.7449, -0.7279, -0.7176, -0.7076, -0.6928, -0.6826, -0.6734, respectively. The CCC value monotonically increases, so it can be concluded that the best *k* according the CCC score is beyond *k* = 1,000, which may not be a useful model for analysis.

Figure 3. BIC score for k = 2, …, 70. The highest score corresponds to k = 24.

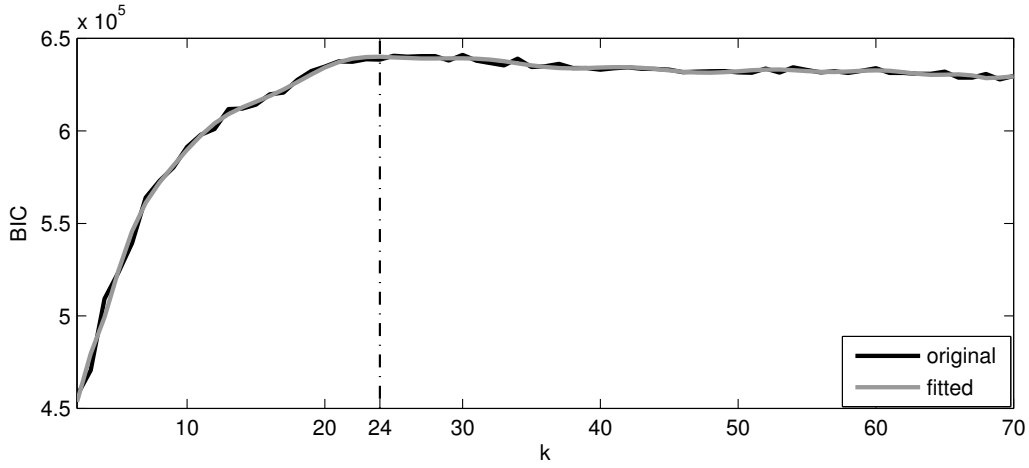


Figure 4. AIC score for k = 2, …, 70. The highest score corresponds to k = 24.

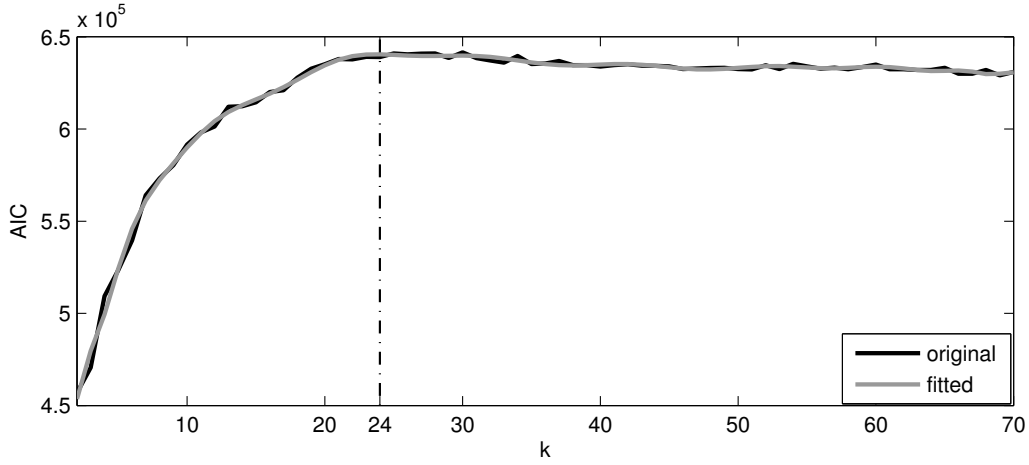


Figure 5. Clustering with k = 24.

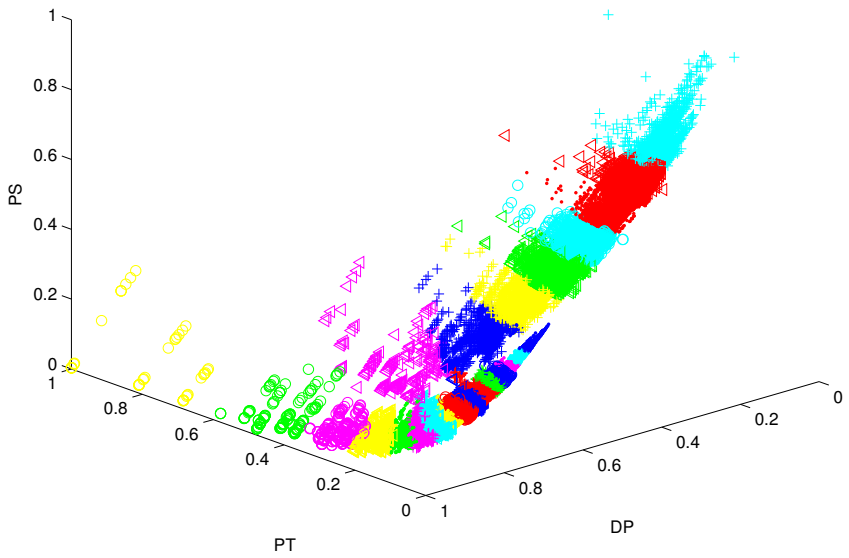
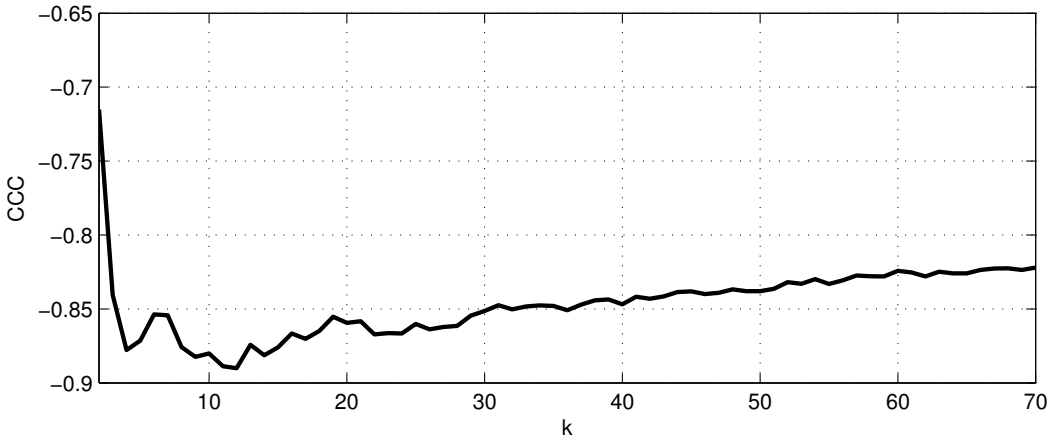


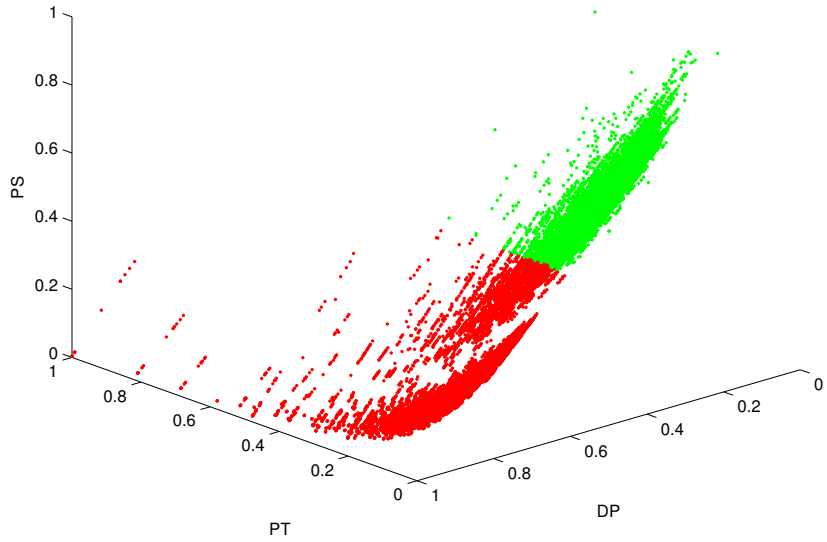
Figure 6. CCC score for k = 2, …, 70.



A similar result is obtained with the ADT. As suggested by Hamerly *et al*. [6], we use $\alpha = 0.0001$ as a confidence level, which corresponds to a critical value 1.8692 for normal distribution ADT. For this critical value, the ADT resulted in 1,908 clusters.

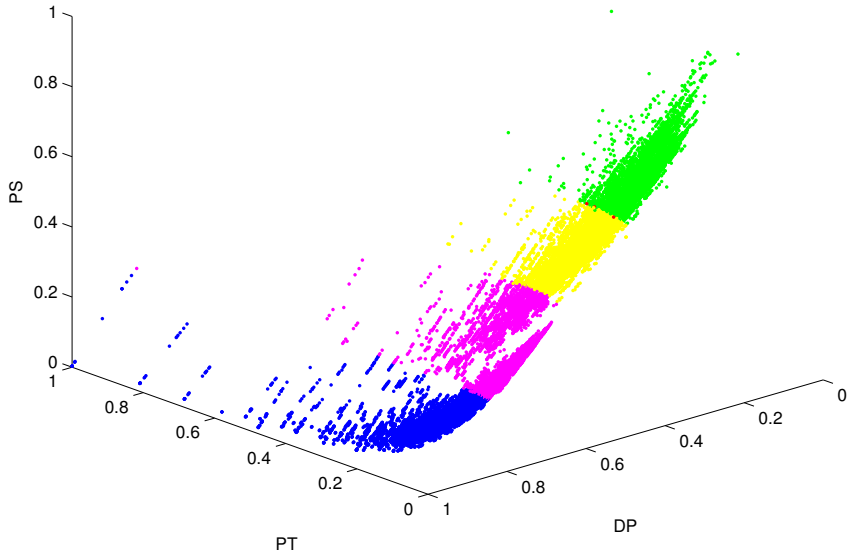
According to the BIC and AIC scores, the best model for the examined data resulted in k=24 clusters. Note that this may be an adequate number of clusters for a 65,520 data points. On the other hand, the CCC and ADT suggest that the number of clusters should be greater than 1,000, which is still too high to be useful for manual analysis. This result may indicate that the data violate the assumption of cluster shape made by k-means, and a clustering method (for example, mean-shift) that can discover clusters with arbitrary shape should be employed. Figures 7 and 8 present the results of applying mean-shift [7] with bandwidth h = 0.5 and the adaptive mean-shift [8] clustering algorithms, respectively.

Figure 7. The result of mean-shift clustering with h = 0.5.



It can be observed that the mean-shift and the adaptive mean-shift result in a lower number of clusters than BIC and AIC. The clustering algorithm can be selected depending on the subsequent analysis. For example, for a detailed examination, smaller clusters discovered by k-means may be more useful, while for a general analysis one may prefer the coarse clustering result produced by the mean-shift. Also note that, the clusters discovered by k-means can be merged to obtain a lower number of clusters.

Figure 8. The result of adaptive mean-shift clustering.



The ideal dataset for *k*-means or mean-shift clustering would consist of scenarios that are in clearly divisible groups which the algorithm can unambiguously identify as separate clusters. For this particular set of data, the cluster boundaries are not obvious, which makes it challenging to interpret. A cluster of interest must be chosen, typically based on an assessment of what properties would be attractive to a proliferator. For example, it is supposed that a proliferator would value a low probability of detection (DP) and a high probability of success (PS). From Fig. 7, the scenarios in the green cluster satisfy both of these criteria and may be a cluster of interest. The cluster is then analyzed to find what inputs tend to lead to a scenario's inclusion in the cluster of interest.

For any given target, the number of scenarios with each of the possible diversion rates is equal over the entire set. This is because every combination of targets and rates was used to create the set. For example, for the TRU Extraction process the proportion of scenarios with 0, 1σ, 2σ, and 4σ are 25% each across the entire set. In the cluster of green scenarios, the proportions are 17%, 26%, 27%, and 31%, respectively. This upward skew indicates that despite a higher likelihood of detection a higher diversion rate at this target tends to lead to a scenario's inclusion in the cluster of interest. The LWR SF Storage target has an opposite skew of diversion rates, suggesting that it is not conducive to a low DP and high PS.

These insights, as well as others based on safeguard conditions, may be applied to efficiently allocate safeguarding resources for planned and existing facilities.

5. CONCLUSION

In this paper, we examine and compare goodness-of-fit measures for automated estimation of number of clusters *k* in the *k*-means clustering algorithm. We observe that these measures tend to favor a higher number of clusters for the data of interest when compared to that of mean-shift and adaptive mean-shift. Noting that the *k*-means algorithm is more time efficient, we suggest that it can be a good alternative to mean-shift when the computational resources are limited.

References

[1] Brookhaven National Laboratory, "*PRCALC Algorithm, Modeling, and User's Guide*," Upton, NY, USA, (2008).

[2] Z. Jankovsky, D. Zamalieva, R. Denning, A. Yilmaz, and T. Aldemir, "*A Comparison of Various Clustering Schemes for Proliferation Resistance Measures*," ANS Winter Meeting, (2013).
[3] R. E. Kass and L. Wasserman, "*A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion*," Journal of the American Statistical Association, 90(431):928–934, (1995).
[4] H. Akaike, "*A new look at the statistical model identification,*" IEEE transactions on automatic control, 19: 716-723, (1974).
[5] P. J. Rousseeuw and L. Kaufman, "*Finding groups in data: An introduction to cluster analysis*," John Wiley & Sons, (1990).
[6] G. Hamerly and C. Elkan, "*Learning the k in k-means*," Neural Inf. Processing Systems, (2003).
[7] K. Fukunaga and L. D. Hostetler, "*The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition*", IEEE Transactions on Information Theory, vol. 21, pp. 32-40 (1975).
[8] B. Georgescu, I. Shimshoni, and P. Meer, "*Mean Shift Based Clustering in High Dimensions: A Texture Classification Example*", Proc. of International Conference on Computer Vision, pp. 456–463, Washington, DC (2003).
[9] M. Yue, L.-Y. Cheng, R. A. Bari, "*A Markov Model Approach to Proliferation-Resistance Assessment of Nuclear Energy Systems*," Nuclear Technology, 162, 26-44 (2008).
[10] D. Pelleg and A. Moore, "*X-means: Extending k-means with efficient estimation of the number of clusters*," Proc. of International Conference on Machine Learning, (2000).
[11] M. A. Stephens, "*EDF statistics for goodness of fit and some comparisons*," American Statistical Association, 69(347):730-737, (1974).