K-Means Clustering Evaluation

Michael Gircys

***Abstract*—The focus of the efforts outlined in this paper were to compare the clustering abilities of the K-means clustering algorithm across various distance measures and centroid counts. A number of synthesized 2D coordinate points were used as sample data, which were measured using the Dunn Index calculation.**

I. Introduction

THE PURPOSE of these experiments was to evaluate the abilities of the k-means clustering algorithm, and the effectiveness of the Dunn Index as a performance measure. Four distance measures are evaluated for each of four data sets to evaluate their effect on Dunn Index performance, where the known optimal K value is used in addition to a number of lesser and greater k values.

The data used for comparison of Dunn Index measures are the S1 to S4 sets from Fränti and Vermajoki [5] and republished by the University of Eastern Finland's Speech and Image Processing Unit. These are synthetic generated points sets with 5000 samples, known centroids, and known variances. In transition from S1 along to S4, we see the known centres of the data points move closer together, which should provide a good variety of group noise, and permits us to better analyse the results when groupings are not linearly separable. The points are in Euclidean space and should be measurable through partitioning on proximity, which is ideal for application with the K-means clustering algorithm.

This report will continue with a review of the K-means clustering algorithm, the distance metrics experimented with, and the Dunn Index calculations considered for performance evaluation. Following the k-means clustering algorithm and equations reviews, we will present and evaluate the clusters found from the execution of the algorithm, and analyse the effectiveness of the Dunn Index. Finally, we will summarize the results found and conclude the experiments.

II. K-Means Algorithm, Review

The K-means clustering algorithm is a quick, partition-based clustering algorithm which can provide the assignment of each given dataset point to one of k groups, and the centroids of each group. K-means has an advantage over some other clustering methods in its speed; the computation and assignment of groupings is done through a relatively quick distance check.

K-means expects both the set of input points, and a provided value k, which specifies the number of discrete groups to which a point may be assigned. On initialization, k centroids are produced either through random selection of the provided data points, or through random generation of each continuous feature for k points. Each data point is then assigned a group based on the centroid with nearest proximity. With the groupings established, the centroids are then updated to take the mean of each feature across all points in its group. The process of assigning groups and updating the centroid is repeated until all points remain in the same group after an iteration. Pseudo-code for the K-means clustering algorithm can be found in Figure 1.

Algorithm 1 K-Means Clustering Algorithm

Require: $K > 0$, and $Points \neq \emptyset$

```
{ Initialize Centroids }
for c = 1 to k do
    Centroids_k ⇐ random point ∈ Points
    Centroids_k.Group ⇐ k
end for

{ Refine Cluster }
GroupingsChanged ⇐ TRUE
while GroupingsChanged = TRUE do
    GroupingsChanged ⇐ FALSE

    { Assign Group }
    for all point ∈ Points = {d_1, ..., d_n, Group} do
        newGroup ⇐ Centroids_i.Group where
            dist(point, Centroids_i) =
            min(dist(point, Centroids_j), 1 ≤ j ≤ K)
        if newGroup ≠ Group then
            GroupingsChanged ⇐ TRUE
        end if
        Group ⇐ newGroup
    end for

    { Update Centroids }
    for all centroid ∈ Centroids = {d_1, ..., d_n, Group}
    do
        Points_i = {p ∈ Points | p.Group = Group}
        for x = 1 to n do
            d_x = mean(Points_i.d_x)
        end for
    end for

end while
```

A. Distance Metrics

The performance of the K-means clustering algorithm will be measured on the same data set over four main distance

metrics. With data in (2D) Euclidean space, We will evaluate distances using the three most notable Minkowski p-norms [3]: 1-norm (Manhattan distance), 2-norm (Euclidean distance), and ∞-norm (Chebyshev distance). Additionally, we will also evaluate Canberra distance, which is a weighted variant of Manhattan distance [1]. While the points evaluated are in (2D) Euclidean space, each distance metric can be generalized for arbitrary dimensions/features. The equations for the evaluated distance metrics are outlined in Figure 1.

Fig. 1: Distance Metric Equations

$$D_{Canberra}(a,b) = \sum_{i=1}^{n} \frac{|a_i - b_i|}{|a_i| + |b_i|}$$

$$D_{Manhattan}(a,b) = (\sum_{i=1}^{n} |a_i - b_i|^1)^1$$

$$D_{Euclidean}(a,b) = (\sum_{i=1}^{n} |a_i - b_i|^2)^{1/2}$$

$$D_{Chebyshev}(a,b) = \lim_{p\to\infty} (\sum_{i=1}^{n} |a_i - b_i|^p)^{1/p}$$
$$= max(|a_i - b_i|, \ldots, |a_n - b_n|)$$

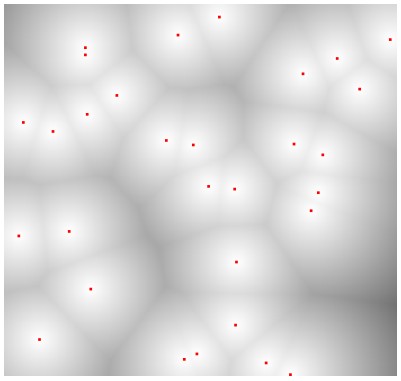
Each of the various distance metrics will obviously have a different bias, and will provide different clustering results for each point on our 2D plane. Generally, Euclidean distance is the most intuitive, and most common measure used for physical measurement. Manhattan distance reflects movement constrained along the axis directions (and is thus aptly named due to its relation to traversal of grid-like Manhattan streets), and Chebyshev distance permits traversal across multiple axis directions simultaneously without loss (such as that count of spaces a King may traverse within the game of Chess). Canberra distance adapts Manhattan distance, where distance is inversely weighted by individual distance of each dimension from the origin. To help visualize some of the bias differences of the distance functions, a type of Voronoi diagram has been provided in Figure 2. The same set of points has been used for a more consistent comparison, and the intensity at each point is a measure of the distance (of the given metric) to the closest red point.

B. Validity Measures

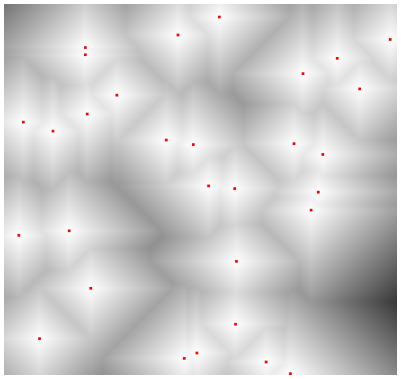
The key measurement that will be explored for clustering performance is the Dunn Index. As shown in the equations in Figure 3, the Dunn Index is a ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. The index provides a result in $[0, \infty]$, with optimal results giving larger values, and indicates that a group of clusters is well separated and compact. Previous experimentation from [7] showed the Dunn Index as having an approximate accuracy of 50% for determining optimal cluster groupings for a specific data set.

Fig. 2: Voronoi Examples of Distance Metrics

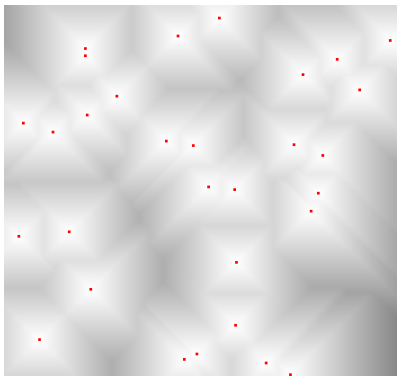
(a) Voronoi: Euclidean Distance



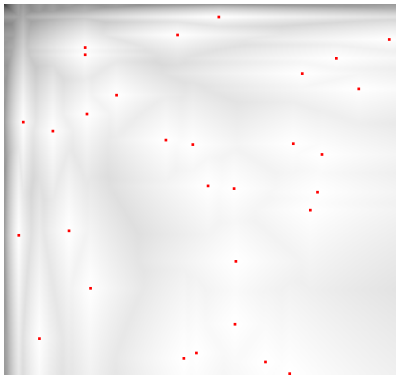
(b) Voronoi: Manhattan Distance



(c) Voronoi: Chebyshev Distance



(d) Voronoi: Canberra Distance



The original Dunn Index measure as used by Dunn [2] will be used. Alternative Dunn-like indices are available, which use various other metrics for the calculation of the intra- and inter-cluster distances. The original measure uses the single linkage measure for inter-cluster distance, though complete, average, and centroid distances would be valid alternatives. For intra-cluster distance, maximum distance between points will be used, though mean distance, and the sum of distance from the mean are also feasible.

Within the experiments performed for this report, the metrics used for the calculation of distance between any two nodes will be the same metric as used within the k-means clustering algorithm.

Fig. 3: Dunn Index Equation

$$\Delta(P_i) = \max_{a,b \in P_i} D(a,b)$$
$$\delta(P_i, P_j) = \min_{a \in P_i, b \in P_j} D(a,b)$$
$$DunnIndex = \frac{\min\limits_{1 \le i < j \le m} \delta(Points_i, Points_j)}{\max\limits_{1 \le k \le m} \Delta(Points_k)}$$

III. Results and Discussion

In measuring the performance of the clustering algorithm with respect to the Dunn Index measure, two variables were adjusted uniformly for each data set. For each of the four data sets, k-means clustering was performed using each of four distance metrics with 5 distinct values for k. Data sets S1, S2, S3, and S4 from Fränti and Vermajoki [5] were used clustered using the optimal k count of 15, and also k values of 9, 12, 18, and 21. Each of the data sets and k values were evaluated over 10 runs using Euclidean, Manhattan, Chebyshev, and Canberra distance metrics. Dunn Index measures used the respective clustering distance metrics of the run which they were evaluating. The random number generator seed was used for all experiments on a given run number to provide consistency with visual comparisons.

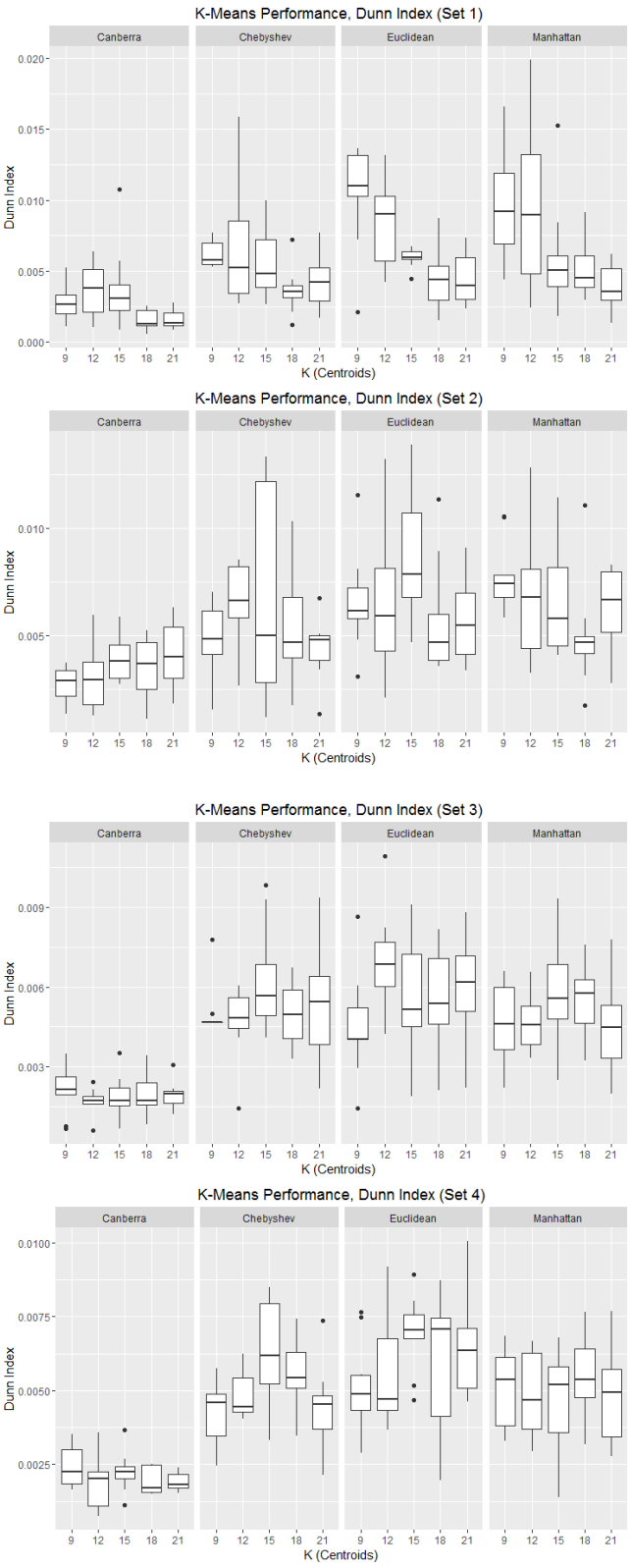
Figure 4 provides a box plot overview of the measured Dunn Index values for each data set, distance metric, and k value.

With a cursory inspection of the produced Dunn Index measures, it is difficult to deduce any clear patterns with k values which are consistent over all data sets. However, a one-way ANOVA across all samples shows a significant correlation (Table I) between chosen k value and Dunn Index, which was further strengthened when the Canberra samples were omitted.

TABLE I: One-Way ANOVA Summary - K-value to Dunn Index

Source	SS	d_f	MS	F	P
Dunn Index	267	1	266.56	15.1	0.0001
Residuals	14133	796	17.71		

Fig. 4: Dunn Index Comparisons



The true optimal k-value appeared to perform best as the data points were grouped closer together, as in the later two sets S1 and S2. In these near proximity sets, performance peaked at $k = 15$, and displayed an somewhat evenly decreasing curve as k positively or negatively diverged. Some curious behaviour can be seen in the results on S1, where smaller k performs much better then larger k, and even better than the expected optimal $k = 15$. The original Dunn Index measure, on these data sets, would appear to more heavily penalize smaller inter-cluster distance, than it does large intra-cluster distance. In sets with greater distinction between known groups, the preference for lower k will show optimal results with many centroids between two or more obvious groupings (as in Figure 7).

Given that the optimal $k = 15$ performed poorly, it is apparent that a number of the centroids were placed too closely, and thus the significance of initial centroid placement should not be overlooked. A number of optimizations for initial centroid placement have been considered by various other authors [6].

One immediate observation about the different distance metrics is the relative small performance measures with the Canberra metric. The distance values produces by the Canberra distance metric were generally much smaller, as the value is weighted by the sum of distance from the origin of the two points. As the data points from the S1 through S4 set were in a $[0, 10^7]$ range, any distance values would be quite small. However, as the Dunn index measure is a ratio of intra- and inter-cluster distance, it would be expected that the scaling of the component distances should not have a substantial effect on this measure. Clearly, however, this distance scheme tends to provide a smaller inter-cluster, or larger intra-cluster distance. This already suggests that there may be potential issues in using a Dunn Index across different distance metrics.

A traditional ANOVA test to evaluate confidence of the effect of distance metric is not suitable here, as the distance metrics are categories, and non-numeric. While Canberra distance generally produced a much lower Dunn Index measure, it should be noted that it was the distance metric responsible for the only fully accurate centroid placement. Canberra excluded, none of the distance metrics appeared to dominate any other metric across all of the data sets. Chebyshev distance appeared to have a higher performance variance with the optimal k-count, and marginally lower overall average performance in most sets, though it appeared to perform slightly better as the set numbers increased, and data points were spaced more closely. Euclidean and Manhattan distances appeared to have similar overall means for this data set. Euclidean distance, except in the case of S1, tended to have better performing outliers, though the statistical significance of these should not be substantial. Generally, the choice of the three main distance metrics did not appear to have substantial bearing on resultant Dunn Index measures, nor on the correct placement of cluster centroids.

The cluster for S1 found to be visually most reflective of the true centroids (Figure 5) had a Dunn Index value of 0.022450103, which was tied for the 30th worst of the 200 runs on that data set. Comparing the best and worst plots of the S1

Fig. 5: S1 Visual Best

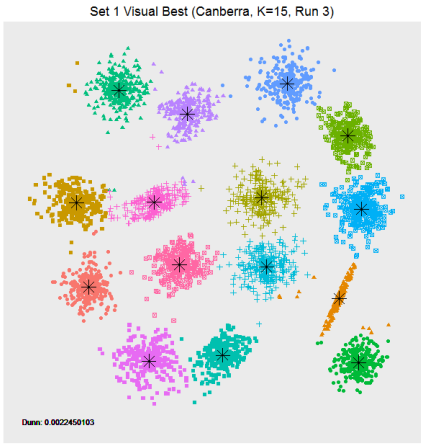


Fig. 6: S1 Worst Dunn Index

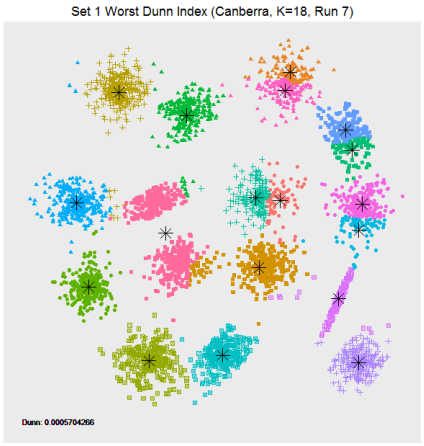
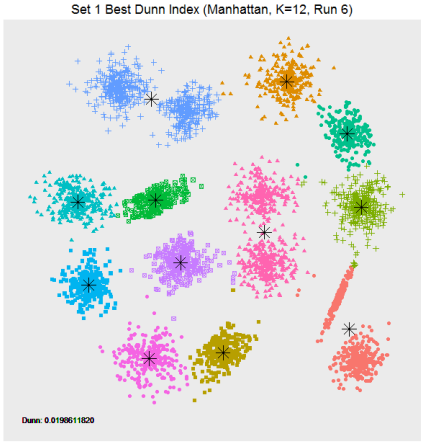


Fig. 7: S1 Best Dunn Index



data set, the graph of the cluster holding the smallest Dunn Index value (Figure 6) does indeed show a poor coherence within partitions. In some areas, visual clusters are shared by multiple centroids, and in others, a single centroid bridges multiple visible clusters. Substantial portions of one group can be owned by a far-off centroid. Yet the plot holding the best Dunn Index (Figure 7) holds examples of these failings as well. Using three centroids shy of the optimal count, there are three centroids responsible for bridging pairs of visual point groups. The visually best plot (Figure 5) has near-perfect placement of centroids, which is the critical, desired result of the k-means algorithm, but is penalized heavily by a small handful of outlier points produced as artifacts from the natural fields about the distance metrics.

Similar results were found with the additional data sets. While the visually worse plots appeared to have substantially lower Dunn Index magnitudes, an overwhelmingly large amount of plots with suboptimal centroids ranked much higher than the plot with accurate centroid placement. As such, it is difficult to make any assertions regarding the necessity of Dunn Index measures for optimal centroid placement or plot aesthetics. With the S1 through S4 data sets, the Dunn Index did not appear to hold any meaningful correlation to better clustering or visuals therein.

IV. Conclusion

While perhaps difficult to initially visualize, a strong correlation appears to exist between chosen k-value and Dunn Index performance. On the S series data sets, the original Dunn Index measure (with single-linkage inter-cluster, and maximum distance intra-cluster measures) on these data sets, would appear to prefer larger, less-concise groupings over smaller groupings that are in close proximity to another. With more distanced points, performance goes to larger k, where values both greater and lesser than optimal k are detrimental in more tightly packed sets.

The Canberra distance metric did frequently produce relatively low Dunn Index measures, but performed sufficiently under visual inspection. With the exception of the lower scoring Canberra distance, the choice of metric did not appear to show substantial correlation with Dunn Index measures, and no correlation was found relating the metrics to aesthetic plot appearance nor correct centroid placement.

While it was seen that plots from clusters with extremely small Dunn Index values were less visually appealing, a number of counterexamples were explored which disproved the converse. A higher Dunn Index value did not correlate with more optimal centroids nor with more aesthetic plots of the points. Rather, small numbers of outlier points, with minor effect on the centroids and plot appeal, appeared to have a heavier than desired influence on the original Dunn Index validation measure. The large amount of plots which ranked higher by Dunn Index than a known optimal solution suggests a weak correlation with Dunn Index and clustering performance, if any, on the data sets explored.

References

[1] Godfrey N Lance and William T Williams. Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*, 1(1):15–20, 1967.

[2] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

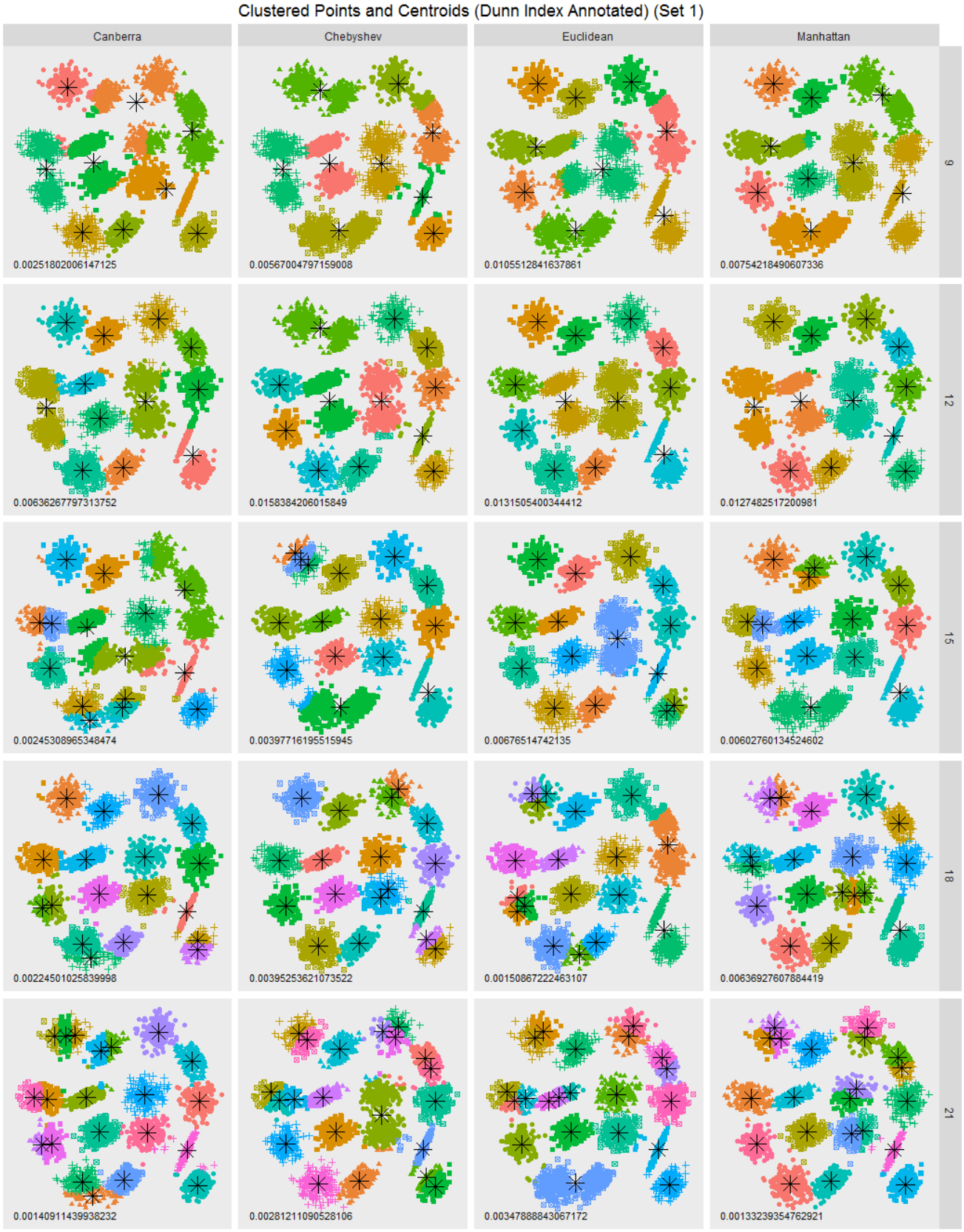
[3] Manabu Ichino and Hiroyuki Yaguchi. Generalized minkowski metrics for mixed feature-type data analysis. *Systems, Man and Cybernetics, IEEE Transactions on*, 24(4):698–708, 1994.

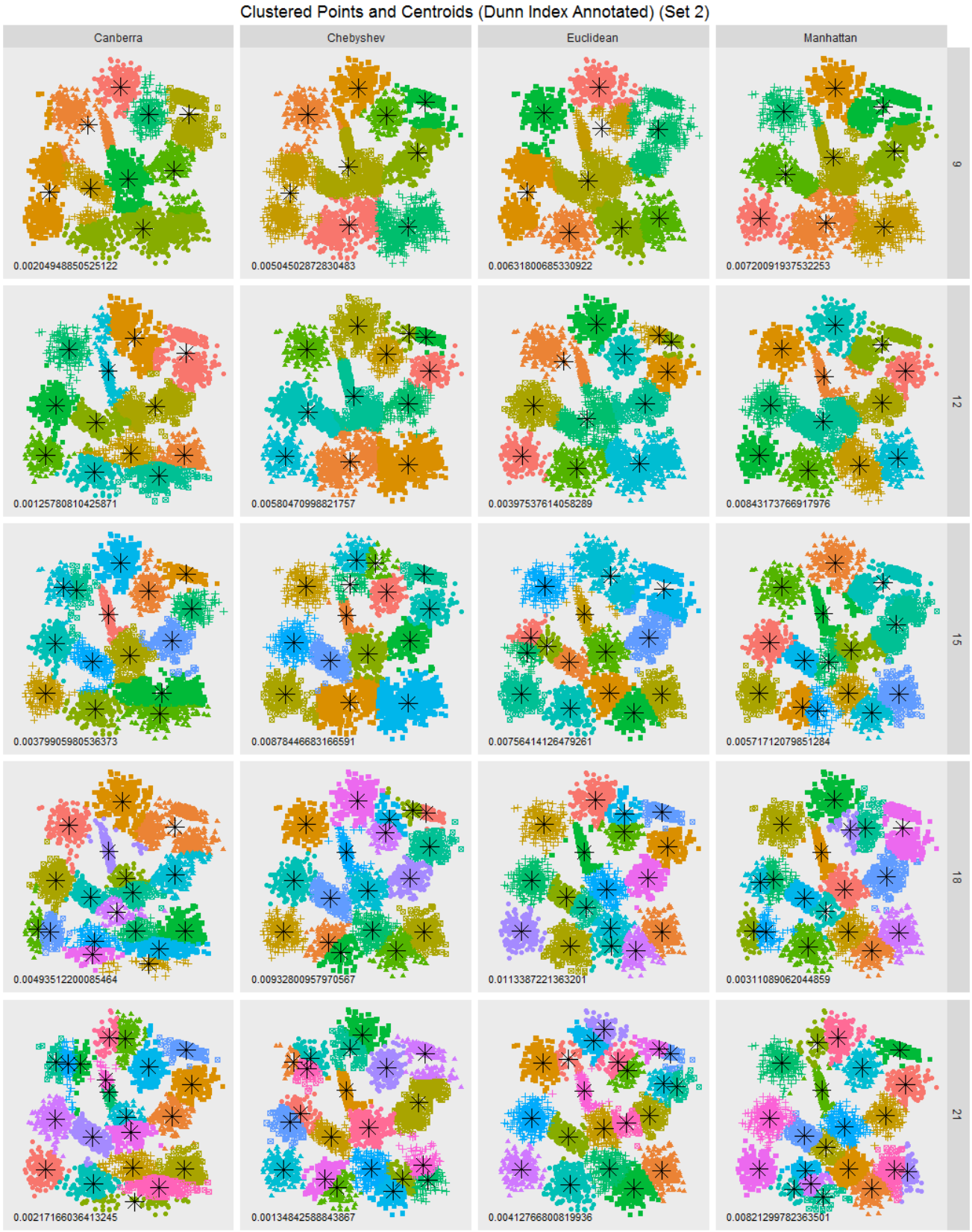
[4] Tom M Mitchell et al. Machine learning, 1997.

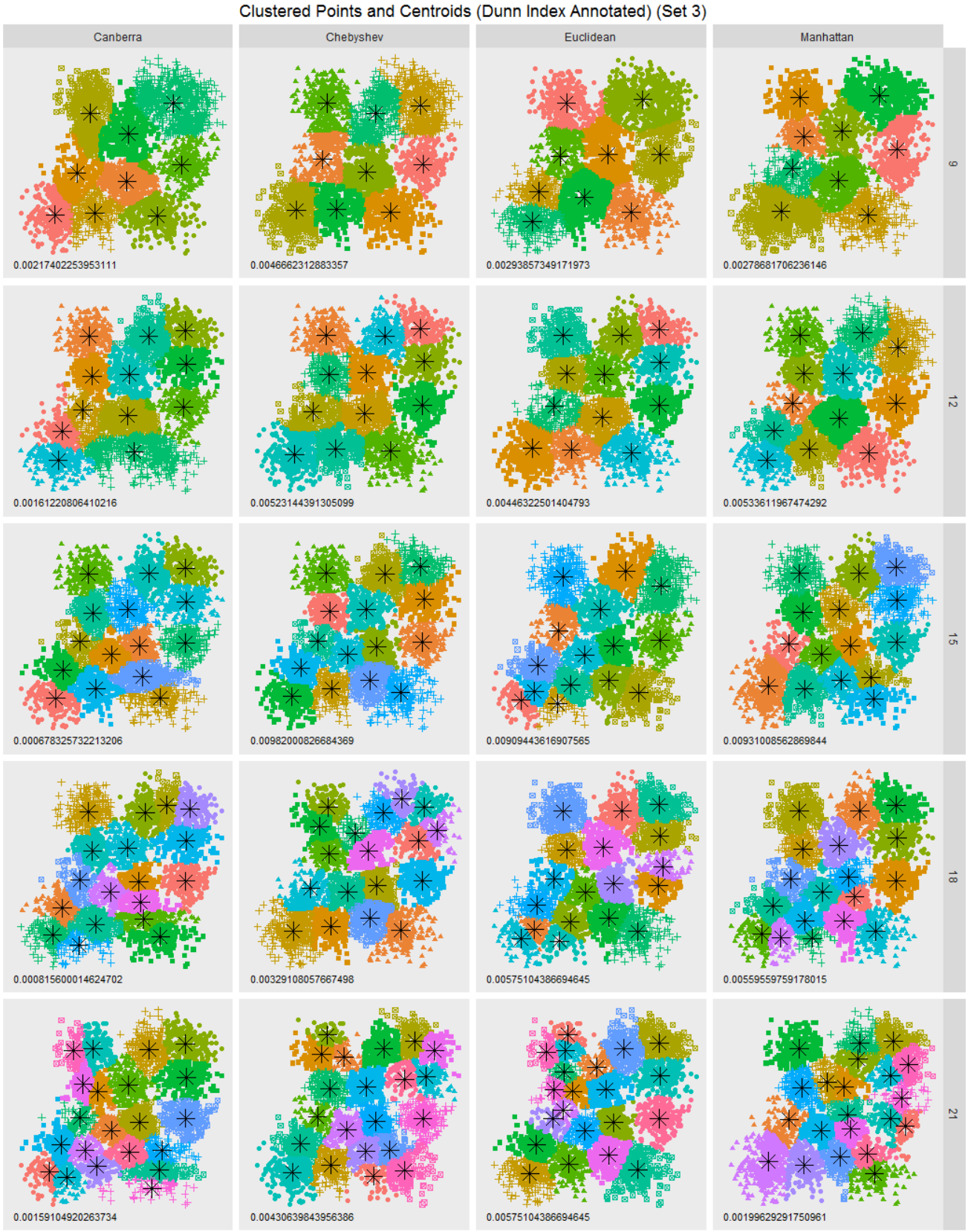
[5] Pasi Fränti and Olli Virmajoki. Iterative shrinking method for clustering problems. *Pattern Recognition*, 39(5):761–775, 2006. [Online; accessed 2016-03-07].

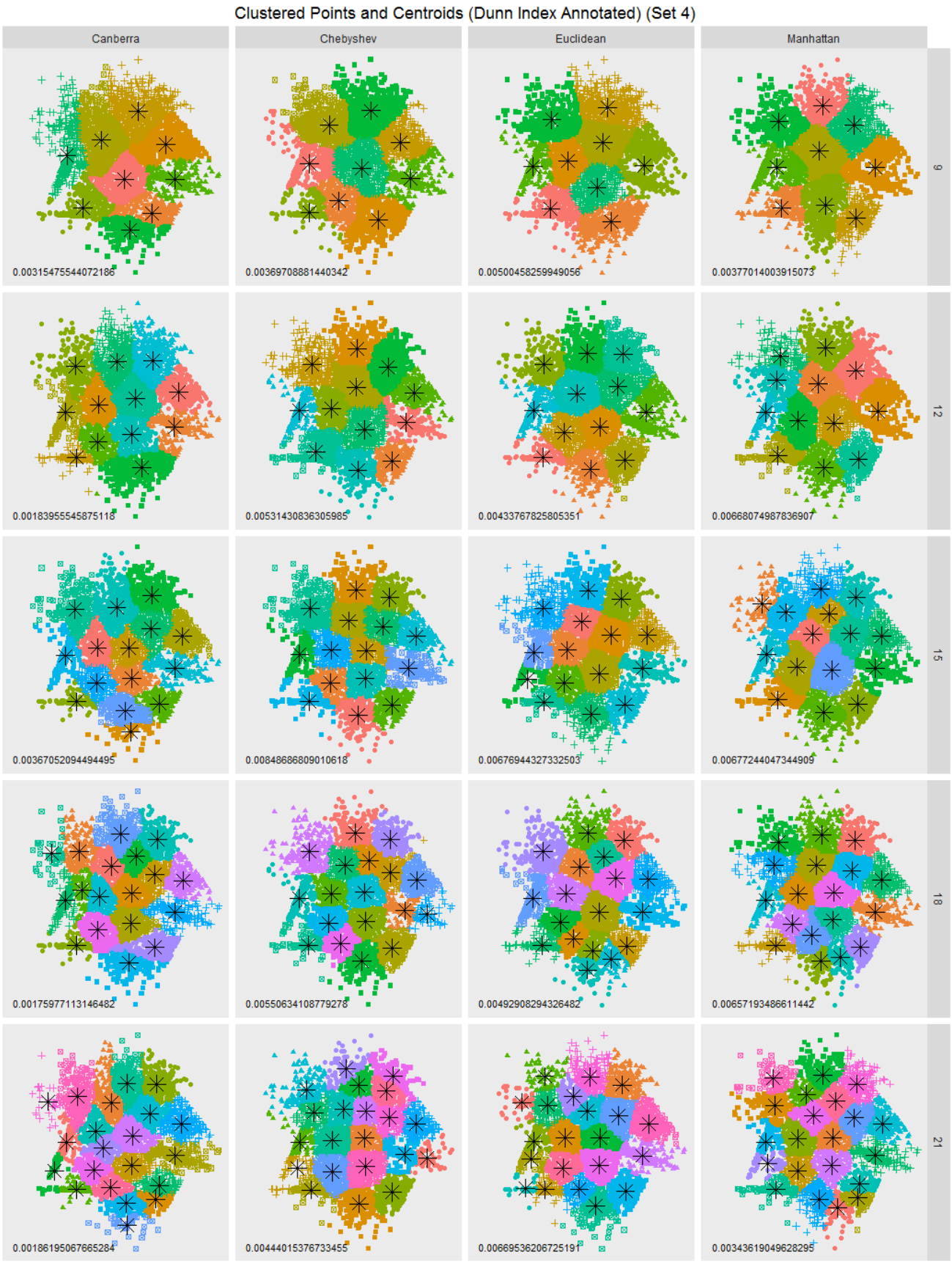
[6] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[7] Unknown. Cluster validity measures. http://www.biomedcentral.com/content/supplementary/1471-2105-9-90-s2.pdf, Unknown. [Online; accessed 2016-03-01].


Clustered Points and Centroids (Dunn Index Annotated) (Set 1)
Canberra
Chebyshev
Euclidean
Manhattan
9
12
15
18
21
0.00251802006147125
0.00567004797159008
0.0105512841637861
0.00754218490607336
0.00636267797313752
0.015838420601584 9
0.0131505400344412
0.0127482517200981
0.00245308965348474
0.00397716195515945
0.00676514742135
0.00602760134524602
0.0022450102583 9998
0.00395253621073522
0.00150867222463107
0.00636927607884419
0.00140911439938232
0.00281211090528106
0.00347888843067172
0.00133239354762921


Clustered Points and Centroids (Dunn Index Annotated) (Set 2)
Canberra
Chebyshev
Euclidean
Manhattan
9
12
15
18
21
0.00204948850525122
0.0050450287283048 3
0.00631800685330922
0.00720091937532253
0.00125780810425871
0.00580470998821757
0.00397537614058289
0.00843173766917976
0.00379905980536373
0.00878446683166591
0.00756414126479261
0.00571712079851284
0.00493512200085464
0.00932800957970567
0.0113387221363201
0.00311089062044859
0.0021716603641324 5
0.00134842588843867
0.00412766800819936
0.00821299782363501


Clustered Points and Centroids (Dunn Index Annotated) (Set 3)
Canberra
Chebyshev
Euclidean
Manhattan
9
12
15
18
21
0.00217402253953111
0.0046662312883357
0.00293857349171973
0.00278681706236146
0.00161220806410216
0.00523144391305099
0.00446322501404793
0.00533611967474292
0.000678325732213206
0.00982000826684369
0.00909443616907565
0.00931008562869844
0.000815600014624702
0.00329108057667498
0.00575104386694645
0.00559559759178015
0.00159104920263734
0.00430639843956386
0.00575104386694645
0.00199629291750961


Clustered Points and Centroids (Dunn Index Annotated) (Set 4)
Canberra
Chebyshev
Euclidean
Manhattan
9
12
15
18
21
0.00315475544072186
0.00369708881440342
0.00500458259949056
0.00377014003915073
0.00183955545875118
0.00531430836305985
0.00433767825805351
0.00668074987836907
0.00367052094494495
0.00848686809010618
0.00676944327332503
0.00677244047344909
0.00175977113146482
0.00550634108779278
0.00492908294326482
0.00657193486611442
0.00186195067665284
0.00444015376733455
0.00669536206725191
0.00343619049628295