# Get a CluE: Ensemble Clustering Chicago Housing Data

David Liao & Cara Bowen-Goldberg

## Abstract

Clustering has been described as in imperfect science and requires a bit of “art”, in terms of domain knowledge and expertise, to help find an optimal number of clusters and validate results[[1]](#endnote-1). Ensemble clustering has not been explored as exhaustively as ensemble classification. This project investigates whether ensemble clustering’s use of parallel soft partitioning methods can improve upon single iterations of certain base clustering algorithms. Appropriate (single) algorithms must be found before attempting to use ensemble clustering. Otherwise, the results can be misleading when comparing the internal cluster quality indices. Our results do show that the BagClust1 consensus criterion with multiple bootstrap replicates of PAM partitioning can refine results of singular PAM partitioning on the same data with the same *k* clusters, and that the results can be validated by internal quality metrics as well as by external validation via visualization (maps) and application of domain knowledge.

## Introduction

### Data: Sources and Pre-Processing

Our dataset is comprised of over 90 attributes for 1,980 geographical tracts in the Chicago area, compiled from the US Census Bureau for the years 2000 and 2013. The attributes contain socioeconomic data including demographics, housing tenure, household income, housing costs, education, population density, age of housing, and housing density. This dataset has been compiled by the DePaul Institute for Housing Studies for a funded study to find naturally occurring housing segments for seven counties. The dataset was split into two subsets of features, one for 2013, and one for ‘Change Over Time’ from 2000 to 2013.

We also use synthetic data, the Cassini dataset, to understand and compare algorithms. The Cassini dataset has two features and three classes, which form two non-convex, banana-shaped clusters curving around a circular cluster between them.

### Motivation and Problem

Initial clustering analyses using k-medoids algorithms and two-step (distance + hierarchical) methods revealed that about k=8 clusters were optimal for both feature subsets. Boxplots and ANOVA tests for the strongest variables were combined with colored-by-cluster geographical maps to check and visualize these results.

However, in a PAM partitioning of the data with 8 clusters, 10-20% of the observations have silhouette widths that suggest they are not well-clustered. The average silhouette width with k=8 is only 0.041, and when we tally up the number of observations with negative silhouette widths of -0.500 to -0.0400, 20% of our dataset is affected. This seems incorrect based on domain knowledge of the diversity of Chicagoland.

### Goal

Using multiple cluster ensemble algorithms and cluster consensus methods, we will investigate whether using ensembles with algorithms besides PAM can match or beat the performance of the single and ensemble PAM partition. We will also check if ensemble clustering with PAM suggests a different optimal *k* clusters, and whether we can improve on the previous partitioning with ensemble clustering. Finally, we will investigate clustering quality validation indices besides silhouette width and external (geographical mapping) validation, in order to better assess the results of our ensemble clustering partitions against our existing single cluster partitions.

## Related Work

Many previous studies have used clustering techniques to analyze census data for the purpose of comparing geographic areas, and looked at demographic, social, and economic attributes that describe the populations and sub-populations of these areas. The quality of clustering results greatly depends on the characteristics of the data being used. Census data, obtained from 5-year rolling samples or decennial full count data, will always be inherently noisy due to the collection process. There is no single clustering method that will work well on all data sets and one should try to match the best algorithm on the particular dataset being studied.

One useful study of clustering with geographic data (Grekousis and Thomas, 2012) studies two variations of fuzzy c clustering, and uses internal cluster quality indices to measure their results.[[2]](#endnote-2) We will be using some of the same internal validation indices in our research to compare k-means with k-medoids and single partitions with ensemble partitions. However, geodemographic segmentation cannot be measured just by internal validation metrics. As this data is about the 7 Counties surrounding metropolitan Chicago, the clustering results should pass external validation. It should make sense when visualized with maps and when examined against the demographic data. Much domain knowledge has already been accumulated by the authors during previous work with the dataset. Single algorithm testing has already been used over several weeks of testing to establish a baseline “truth” that when using PAM, k=8 have represented the 7 Counties of Chicago well.

## Methodology

All procedures are performed on each of our two subsets of data (the “now” aka 2013 data, and the “change” (over time) data), as well as on a synthetic dataset (see Appendix A, Section 3). First, we run simple single clustering of three k-means algorithms (Hartigan-Wong, “HW”; Lloyd aka Forgy, “Lld”; and MacQueen, “MQ”). We also run three different single clusterings using PAM, each with a different *k* clusters. (Hierarchical clustering was outside our realm of interest for this report; density-based clustering was explored using a synthetic dataset, but was found to be inappropriate for the Chicago data. See Appendix A, Section 1).

Working with R and the CLUE package[[3]](#endnote-3), we then create ensembles of:

* bootstrapped samples with PAM, varying *k* (7, 8, and 9) and number of replicates (50 and 100)
* different combinations of the three variants of kmeans partitions
* bootstrapped samples with kmeans, varying the algorithm (HW, Lld, and MQ) and number of replicates (50 and 100)

We then generate a ‘cluster consensus’ for each PAM partitioning using the BagClust1 algorithm[[4]](#endnote-4). And for the kmeans partitionings, we use generate three different consensuses for each: BagClust1, SE (soft Euclidean)[[5]](#endnote-5) and GV3 (Gordon and Vichi’s “third model”)[[6]](#endnote-6) (see Appendix A, Section 2). With all of our ensemble partition consensuses formed, we explore appropriate internal validation metrics (cluster quality indexes) to each consensus. Finally, we compare these with external validation (geographic maps by cluster and by variable) and apply domain knowledge to decide if our results make sense.

## Results

### A Look at “Weak Membership”

Each consensus object in the R environment contains a membership matrix, which represents the final partition that a given consensus method has ‘voted on’, given the multiple bootstrapped partitions in a given ensemble. As our Census Tract dataset contains just under 2000 observations, bootstrap sampling should provide the diversity of input data for our ensemble runs. The membership matrix shows how strongly each data row ‘belongs’ to a particular cluster, showing the proportion (between 0 and 1) of votes it received. For this project, when a given observation’s proportions are all found to be less than 0.5, we will describe this observation as being a ‘weak member’ (of all clusters) in the consensus partition, since this observation is getting a split vote and not a majority win. We will use this ‘weak membership’ criteria to count the total observations failing to get a stable cluster assignment from each ensemble-consensus combination.

This is not necessarily a measure of quality of a given partitioning method, but can help us evaluate a cluster’s quality. This is similar to the silhouette width measure that can be obtained from PAM, as we discussed in the introduction. Silhouette widths of each observation can be examined, and values closer to 1 indicate a stronger likelihood of that observation belonging to that cluster. When the silhouette widths for all observations in a cluster fall within a smaller range, this generally indicates more cohesion in that cluster. Still, there will always be observations that rightly fall somewhere in the ‘gaps between’ clusters, similar to a support vector in SVM, and forcing a majority vote upon such an observation does not mean that observation has received a ‘good’ cluster assignment – and indeed forcing iffy observations into a given cluster can reduce that cluster’s quality.

### The Optimal Number of Clusters for Ensemble PAM Remains k=8

As we varied *k* from 7 to 9 for both datasets, we obtained the least number of rows of “weak” members when k=8 (see Figure 1). Partitioning into 8 clusters does indeed seem to suit both datasets allowing for more rows to be definitively assigned by majority vote.

Figure 1: Ensemble with 100 Bootstrap Replicates of PAM Clustering + BagClust1 Consensus

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DATASET** | ***k*** | **CLUSTER SIZES** | | | | | | | | | **WEAK MEMBERS** |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Change | 7 | 550 | 425 | 206 | 121 | 484 | 74 | 46 | --- | --- | 244 |
| 8 | 374 | 380 | 164 | 292 | 116 | 457 | 93 | 28 | --- | 196 |
| 9 | 351 | 39 | 148 | 108 | 290 | 112 | 291 | 439 | 28 | 290 |
|  |  |  |  |  |  |  |  |  |  |  |  |
| Now | 7 | 292 | 54 | 252 | 461 | 300 | 246 | 304 | --- | --- | 190 |
| 8 | 269 | 154 | 159 | 451 | 295 | 221 | 215 | 145 | --- | 167 |
| 9 | 214 | 219 | 175 | 136 | 342 | 276 | 193 | 211 | 143 | 237 |

We can also see in Figure 1 that when partitioning the Change data into seven clusters, resulting clusters 1 and 5 are both large in size, with 550 and 484 members respectively. Given there are 1906 observations in the Change data set, two clusters containing 55% of overall members does not jibe with Chicagoland’s diversity. This is based on many weeks of domain knowledge acquired in regards to the complexity and socioeconomic diversity of the 7 Counties area of Chicago. Cluster sizes were one of the evaluation methods used in the research project and correct cluster sizes came from examining geographic maps beside boxplots of all variables against cluster groups, in order to find dominant variables that explain each cluster’s characteristics.

### Can Clustering Ensembles Improve PAM Clustering of Chicago Housing Data?

Now that we have confirmed 8 is the best number of clusters, we continue with the procedures we described in the Methodology section. The results table is presented in Figure 2. We see that PAM with no ensemble on the Change data returns 213 observations that, according to their silhouette widths, are classified as “weak members”, while PAM with no ensemble on the NOW dataset returns 220 such observations. But when we create ensembles of 50 and then of 100 bootstrap replicates of PAM on the Change data, and extract a consensus using BagClust1, those “weak members” are reduced from 213 (no ensemble) to 203 (with 50 replicates) to 199 (with 100 replicates); and the same method on the Now data reduces weak members from 220 (no ensemble) to 158 (50 replicates) to 129 (100 replicates). We suspect the stronger improvement with the Now dataset is because the absolute values of variables in the Change dataset are sometimes very small and quite close to zero. This makes the Change dataset inherently more difficult to cluster, as there is less variability amongst the 34 variables.

Amongst the kmeans ensembles of the Now data, we see the consensus criterion GV3 produced far fewer weak members than SE. The kmeans ensembles of the Change data with the consensus criteria GV3 and SE produce almost no weak members, but with consensus criterion BagClust1 we see that Hartigan-Wong (known as the ‘smartest’ of the three kmeans methods) has the fewest weak members.

#### Evaluating Cluster Quality: Internal Validation Indices

We seek to maximize Dunn Index, Calinski-Harabasz, and Silhouette; we seek to minimize Davies-Bouldin, Xie-Beni, and SD Index[[7]](#endnote-7). On the indices we seek to maximize, on both datasets, PAM with 100 bootstrap replicates and BagClust1 consensus criterion outperforms bagged PAM with 50 replicates and PAM with no ensemble (see Figure 2). On the indices we seek to minimize, on both datasets, PAM with 100 bootstrap replicates and BagClust1 outperforms the other two PAM models on Xie-Beni. For the 2013 data, it also performs best on Davies-Bouldin.

Figure 2: Results Table. Cluster Sizes, Weak Members, and Internal Cluster Quality Indexes.

These internal clustering validation indices are, at best, measures relative to each other, and behave differently on different types of datasets. They cannot be used alone as arbiter of what algorithm gives best clustering quality. This is the art and science of clustering where domain knowledge, cluster sizes, visualization and validation indices must all be evaluated together as relative measures to each other to help consolidate conclusions.

#### External Validation

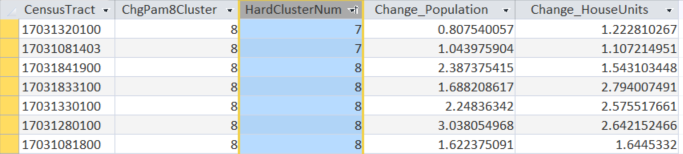
While the intent of this project is not to describe the characteristics of a given set of clusters, it is useful to ‘label’ cluster in the Change data for purposes of external validation (see Figure 3 for cluster descriptions; see Appendix B for maps shaded by these clusters and accompanying discussion).

Figure 3 Labels for clusters from partitioning Change dataset with k=8

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| C 1 | C 2 | C 3 | C 4 | C 5 | C 6 | C 7 | C 8 |
| Little change | Hardship, inner city | Young, gentrification | Hardship, outer city | Wealthy, mostly suburban | Middle class, suburban | Medium growth\* | High growth\* |

With the bagged PAM ensemble, the High Growth cluster (#8) has 27 tracts, versus 50 tracts using PAM with no ensemble. The High Growth cluster is mostly characterized by large positive changes in Population and House Units Density. An examination of some sample tracts re-assignment shows some “weak members” for these two variables were indeed aptly re-assigned by the PAM ensemble (see Figure 4).

Figure 4 Change in membership of Cluster 8, single PAM vs PAM ensemble, Change dataset

****

We characterize cluster 3 as the Gentrification cluster as it has the largest increases in white population and in people with a bachelor’s or higher degree, also increases in housing cost burden greater than 30% of household income. It’s also the only cluster to have decreases in Hispanic population and Race=”White Alone”, Education Attainment=”Bachelors Degree or More”, and increases in Cost Burden=”Greater Than 30% of Household Income”. It’s also the only cluster to have decreases in Hispanic population. The reassignment of tracts between the single PAM clustering and the bagged PAM clustering seems justified when we compare peer tracts for these variables (see Appendix B).

For the Change data, the PAM ensemble does appear to refine PAM without ensemble, and this refinement makes sense to our domain knowledge. Upon careful examination, various tracts do seem to have been re-assigned from one cluster to another appropriately.

### Applying Ensemble Methods to the Wrong Base-Clustering Algorithm Does NOT Improve Results

The kmeans algorithm is known to be sensitive to outliers and it was already determined by the research project that it was likely not suitable for housing segmentation. Boxplots of the eight clusters resulting from PAM clustering easily show that some variables have quite a few outlier spikes (see Appendix B). Still, we apply ensemble kmeans to our noisy data to see if the consensus voting approach could at all overcome the weaknesses of kmeans regarding outliers. The first sign that ensembles can’t improve on the wrong base algorithm is the resulting cluster sizes. Every kmeans ensemble returns one cluster with only 8 or 9 members, and another tract with over 600 (see Figure 2). It is also suspicious that only 1 or 0 tracts are classified as “weak members”. This suggests that kmeans, even in an ensemble, is cutting through the decision space too bluntly.

We can also learn from these results about the need to backup internal validation with external validation, since the kmeans algorithms achieve the “best scores” overall across validation measures on both datasets, but these scores are meaningless when checked against our domain knowledge. The ‘best’ kmeans ensemble, according to the Davies-Bouldin index, uses MacQueen’s kmeans method and the GV3 consensus criterion. We focus in on a “Wealth” cluster with common census tracts in both the 100-bagged PAM partition and this kmeans ensemble (see Appendix C, Figure 1.1). The PAM ensemble has assigned 119 tracts to this cluster, while this ‘best’ kmeans ensemble has assigned it only 9 tracts. This is simply incorrect when we look at the distributions of the attributes Median Household Income and Median Home Value. See Appendix C for more comparisons of PAM to kmeans ensembles, all of which show inaccurate assignments of tracts to clusters when using kmeans.

## Conclusion

Applying an inappropriate algorithm to a given dataset can return misleading results that cannot be corrected by ensemble clustering, regardless of consensus method or other parameters. This was demonstrated in both the toy dataset and the Chicago Housing dataset. Otherwise, one can easily be fooled by measures of clustering quality from very low “weak members” in the membership tables of ensemble clustering or any of the six validation indices we tested.

Running greater numbers of parallel partitions before the consensus voting does reduce the number of weakly clustered members and improve the validation indices scores. For each dataset, there is likely a range of minimum parallel partitions to be used in ensemble clustering, where the clustering results stabilize and converge and would be reflected via internal validation.

The resulting “weak members” after an ensemble clustering run can help in finding the optimal *k* clusters. This should not be used on its own but as a complement to other traditional techniques of finding *k*. The membership matrix can be seen as a snapshot of how confident that ensemble results are in voting each row into the best cluster.

External validation was quite important in checking if ensemble clustering does correctly improve the re-assignment of a census tract from one cluster in PAM with no ensemble to another in PAM with ensemble. We cannot assume that the influence of diversity injected by bootstrap replicate sampling and consensus voting will always improve the clustering assignments of each census tract. Without a systematic numeric method to find and check each tract re-assigned from our best PAM ensemble, we spot-checked various clusters’ re-assigned tracts and confirmed the logic of re-assignment, using the most influential attributes in each cluster as well as geographical maps colored by clusters from each partitioning.

The six internal validation indices we looked at should be used carefully. The best scores for each individual indices for any particular clustering algorithm with a chosen consensus voting method may not always indicate the true best performer. All six scores or more should be used together to help evaluate relative performance of each algorithm. Using an inappropriate validation index, such as a density-based index against a dataset not appropriate for density clustering, will mislead.

## References

1. R. Ghaemi, M. N. Sulaiman, H. Ibrahim, N. Mustapha "A Survey: Clustering Ensembles Techniques", World Academy of Science, Engineering and Technology, vol. 26, pp. 636-645, 2009. [↑](#endnote-ref-1)
2. G. Grekousis, H. Thomas, “Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The Fuzzy C-Means and Gustafson–Kessel methods”, Applied Geography, vol. 34, pp. 125-136, May, 2012. [↑](#endnote-ref-2)
3. K. Hornik, clue: Cluster ensembles. R package version 0.3-51., 2016; K. Hornik, “A CLUE for CLUster Ensembles.” Journal of Statistical Software, vol. 14, no. 12, 2005. [↑](#endnote-ref-3)
4. S. Dudoit, J. Fridlyand, “Bagging to improve the accuracy of a clustering procedure, Bioinformatics, vol. 19, no.9, pp. 1090-1099, 2003. [↑](#endnote-ref-4)
5. K. Hornik, W. B¨ohm, “Hard and Soft Euclidean Consensus Partitions”, Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft f¨ur Klassifikation e.V., Albert-LudwigsUniversit¨at Freiburg, pp. 147–154, 2008. [↑](#endnote-ref-5)
6. A.D. Gordon, M. Vichi, "Fuzzy partition models for fitting a set of partitions", Psychometrika, vol. 66, no. 2, pp 229-247, , June, 2001. [↑](#endnote-ref-6)
7. Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, "Understanding of Internal Clustering Validation Measures", Proceedings of the 2010 IEEE International Conference on Data Mining, pp 911-916, 2010. [↑](#endnote-ref-7)