NYC Restaurant Inspections

Analysis and k-Means Clustering

Andrew Kaluzny

November 2, 2014

Sources

- MacQueen, J. (1967) "Some Methods for Classification and Analysis of Multivariate Observations." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds L. M. Le Cam and J. Neyman. Berkeley, CA: University of California Press.
- (2014) "Determining the number of clusters in a data set." Wikipedia. Wikimedia Foundation, Inc. `http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set`
- (2014) "k-means clustering." Wikipedia. Wikimedia Foundation, Inc. `http://en.wikipedia.org/wiki/K-means_clustering`
- R Core Team. (2014) "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing. Vienna, Austria. `http://www.R-project.org`

Sources: Data

- (2014) "DOHMH New York City Restaurant Inspection Results." NYC Open Data. New York, NY: The City of New York. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/xx67-kt59
- (2012) "How We Score and Grade." New York City Department of Health and Mental Hygiene. New York, NY: The City of New York. http://www.nyc.gov/html/doh/downloads/pdf/rii/how-we-score-grade.pdf
- (2010) "Self-Inspection Worksheet for Food Service Establishments." Bureau of Food Safety and Community Sanitation. New York, NY: The City of New York. http://www.nyc.gov/html/doh/downloads/pdf/rii/self-inspection-worksheet.pdf

Restaurant Inspections

- Began July 2010
- Itemized violations contribute to a score based on severity
  - 2B: Hot food not held at or above 140°F, 7 to 28 points
  - 10J: "Wash Hands" sign not posted at hand-wash facility, 2 points
  - etc.
- A: 0–13, B: 14–27, C: 28 and higher
- Not all inspections are graded, low grades lead to re-inspection

Getting and Cleaning the Data

- Data set available through NYC Open Data
- Data needs to be cleaned
  - e.g. `Fontana"s` $\rightarrow$ `Fontana's`
- Data needs to be parsed
- Code performance concerns
  - $\approx 24,500$ rows of data
  - R quirks

Preliminary Analysis

Time between inspections

- Mean time between inspections around 130 days
- Inspections that end in an A grade have a shorter time since last inspection (about 120 days)

Number of inspections

- Mean of about 7 inspections per restaurant

Score

- Mean score of 16
- time averaged to 12

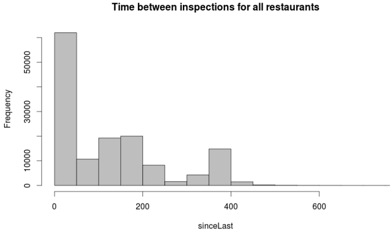
Brough Differences

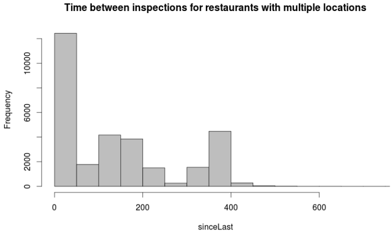
- Brooklyn and Queens have higher avg. scores
- Staten Island has lower avg, but higher time-avg.
- Relative frequencies of cuisines

Multiple Location

Restaurants with multiple locations

- Lower scores
- Longer times between inspections
- Pancakes/Waffles, Donuts, Hamburgers, Sandwiches

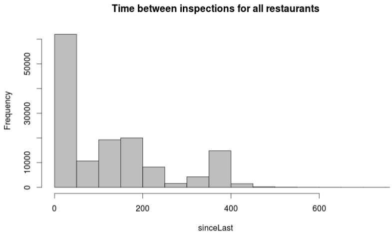


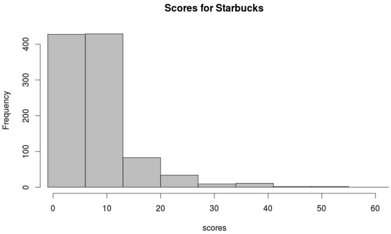


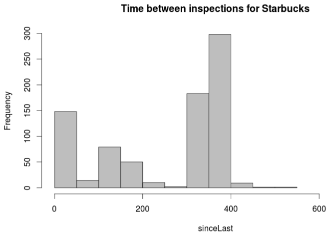
Where are Restaurants with Multiple Locations?

Starbucks

- 229 locations
- avg. score 8.7, time-avg. 6.9
- avg. 250 days between inspections
- avg. 4.5 inspections







Clustering

Why do we cluster things?

- Exploratory analysis
- Classification

k-Means clustering

- Clusters defined by their means
- Originated as an information theory problem (S. P. Lloyd, 1957)
- Analogy to the case of estimating a single mean

k-Means Clustering

Notation

Event space E

Probability mass function p

$\{z_i\}_{i=1}^{\infty}$ random points in E

$x = \{x_i\}_{i=1}^{k}$, $x_i \in E$

Partition $S = \{S_i\}_{i=1}^{k}$

$$\mu_i = \frac{\int_{S_i} z \, dp(z)}{p(S_i)}$$

Given x, define $S(x) = \{S_i(x)\}_{i=1}^{k}$ the minimum distance partition of E

Algorithm (MacQueen)

At each step n we have the k-means $x^n = \{x_i^n\}_{i=1}^k$, (integer) weights $w^n = \{w_i^n\}_{i=1}^k$, and partition $S^n = S(x^n)$

At the start

$$x_i^1 = z_i \qquad\qquad w_i^1 = 1$$

For each subsequent step, we incorporate a new point z_{k+n} and update

$$\text{if } z_{k+n} \in S_i^n \text{ then } x_i^{n+1} = \frac{x_i^n w_i^n + z_{k+n}}{w_i^n + 1}$$
$$w_i^{n+1} = w_i^n + 1$$
$$x_j^{n+1} = x_j^n \text{ and } w_j^{n+1} = w_j^n \text{ for } j \neq i$$

Pseudocode Algorithm

```
x[1:k] <- z[1:k]
w[1:k] <- 1
for i in 1:n {
    find j that minimizes distance(x[j], z[i+k])
    x[j] <- (x[j] * w[j] + z[i+k]) / (w[i] + 1)
    w[i] <- w[i] + 1
}
```

Convergence of k-Means

We define

$$W(x) = \sum_{i=1}^{k} \int_{S_i} |z - x_i| dp(z)$$

$$V(x) = \sum_{i=1}^{k} \int_{S_i} |z - \mu_i(x)| dp(z)$$

Theorem

The sequence $\{W(x^1), W(x^2), \ldots\}$ *of random variables converges and* $\lim_{n\to\infty} W(x^n) = V(x)$ *for some* x *where* $x_i = \mu_i$ *and* $x_i \neq x_j$ *for* $i \neq j$.

A sketch of the proof

A Helpful Lemma

Lemma

For sequences of random variables $t_1, t_2, \ldots$ and $\epsilon_1, \epsilon_2, \ldots$ with a monotone increasing set of events $\beta_1, \beta_2, \ldots$, if

$$|t_n| \leq K < \infty$$

$$\epsilon_n \geq 0, \sum_{n=0}^{\infty} \epsilon_n < \infty$$

$$E(t_{n+1}|\beta_n) \leq t_n + \epsilon_n$$

Then $t_1, t_2, \ldots$ and $s_0, s_1, \ldots$ converge, where $s_0 = 0$ and $s_n = \sum_{i=1}^{n}(t_i - E(t_{i+1}|\beta_i))$

Pathological Distributions

- Circle
- Square
- Rectangle

What do we want to cluster?

Frequency of violations

- Tally occurrences of each violation
- Scale by number of inspections

Transitions between grades

- Treat the grades as a Markov chain
- Build a matrix of transition probabilities
- How to treat transitions we don't have data for?

Tabulating Violation Frequencies

```
for id in restaurantIDs {
    get violations for id
    for v in violations {
        result[v, id]++
    }
    scale result[, id] by number of inspections
}
result <- transpose(result)
```

Creating Transition Matrices

```
for id in restaurantIDs {
    get grades for id
    create empty 3 by 3 transMatrix
    for ii in 2:n {
        lastGrade <- grades[ii-1]
        curGrade <- grades[ii]
        increment transMatrix[curGrade, lastGrade]
    }
    for ii in 1:3 {
        scale transMatrix[, ii] by sum of column
    }
    result[, id] <- transMatrix
}
result <- transpose(result)
```

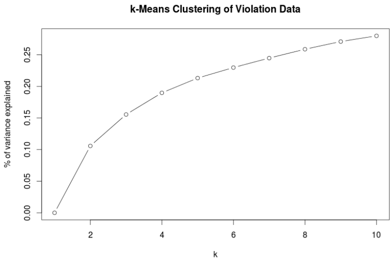
Finding k

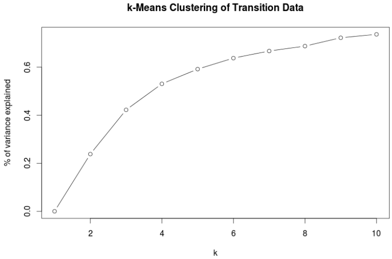
How do we find k?

- Increasing k will always reduce sum-of-squares

The elbow method

- Find where increasing k has less of an impact (the "elbow")





Clustering Results

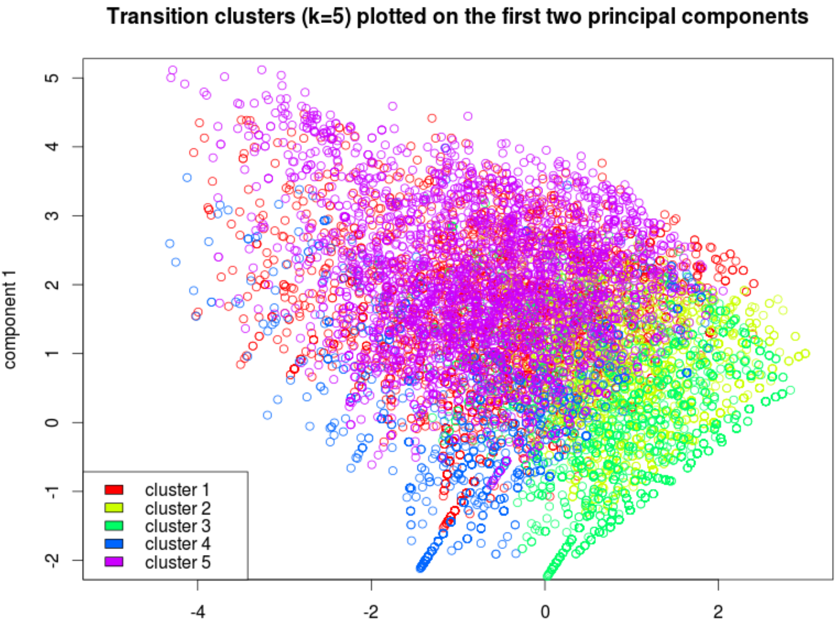
Violation frequencies

- Very sparse data
- Clusters don't account for much of the variance (about 25%)
- Hard to interpret

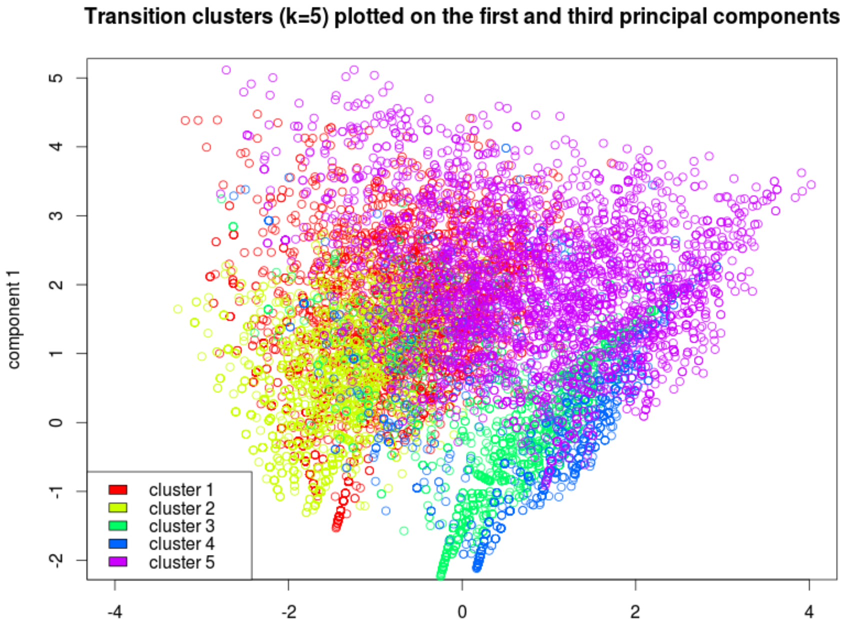
Transition matrices

- Good accounting of the variance (about 60%)
- Can interpret cluster centers
- Problems of scaling from missing data

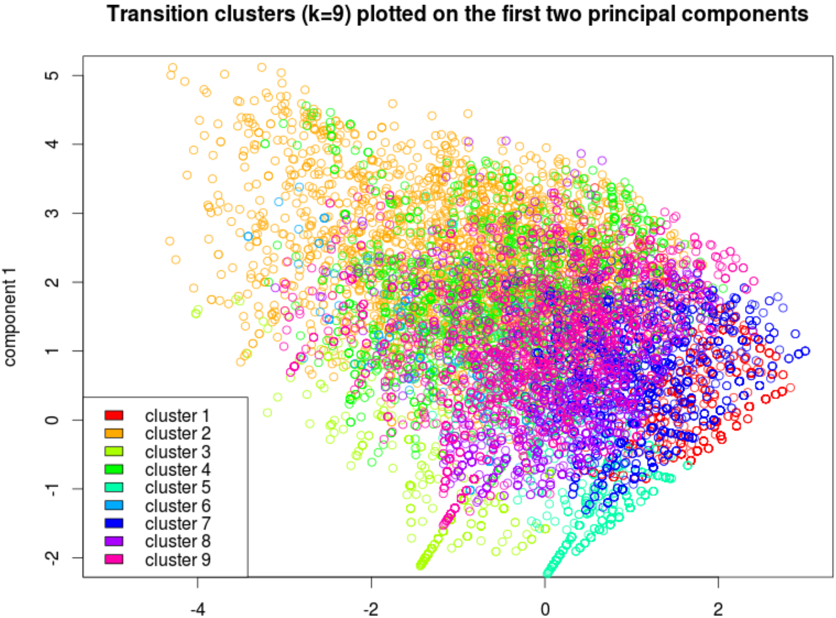
Visualizing Clusters



Visualizing Clusters



Visualizing Clusters



Transition Clusters

$k = 5$

- $A \to A$, $B \to B$, and $C \to C$ dominated restaurants all end up clustered together
- Two clusters with dominant $B \to A$ transitions
  - one has notable $A \to A$ and $A \to B$ transitions, with barely any transitions to C (#1)
  - other splits evenly from A, when at C, the $C \to A$ transition is dominant

Transition Clusters

$k = 9$

- $A \to A$ dominated cluster is identifiable
- "re-scaling" of matrices helps make sense of centers
- Clusters with dominant $B \to A$ transitions still identifiable
  - Third one appears that drifts down, with relatively small $A \to A$ transition
- Sizeable cluster with a strong $C \to A$ transition that then goes between A and B with slight chances of dropping to C
  - Looking at unscaled version, see most of the data comes from transitions out of C

Conclusion

Conclusions!

Questions?