A CLUSTERING APPROACH FOR US HISPANIC HOUSEHOLDS SEGMENTATION

Juan Sebastián Marín-Delgado

Tutor:
Francisco Zuluaga-Díaz

Proposal Presentation
February 25th, 2015

But before starting…

Question:

What would you answer if you were asked on how to target a specific group of a given population?

On the other hand:

If you are given a very large and heterogeneous population, how can you group the population into less heterogeneous subpopulations?

Agenda

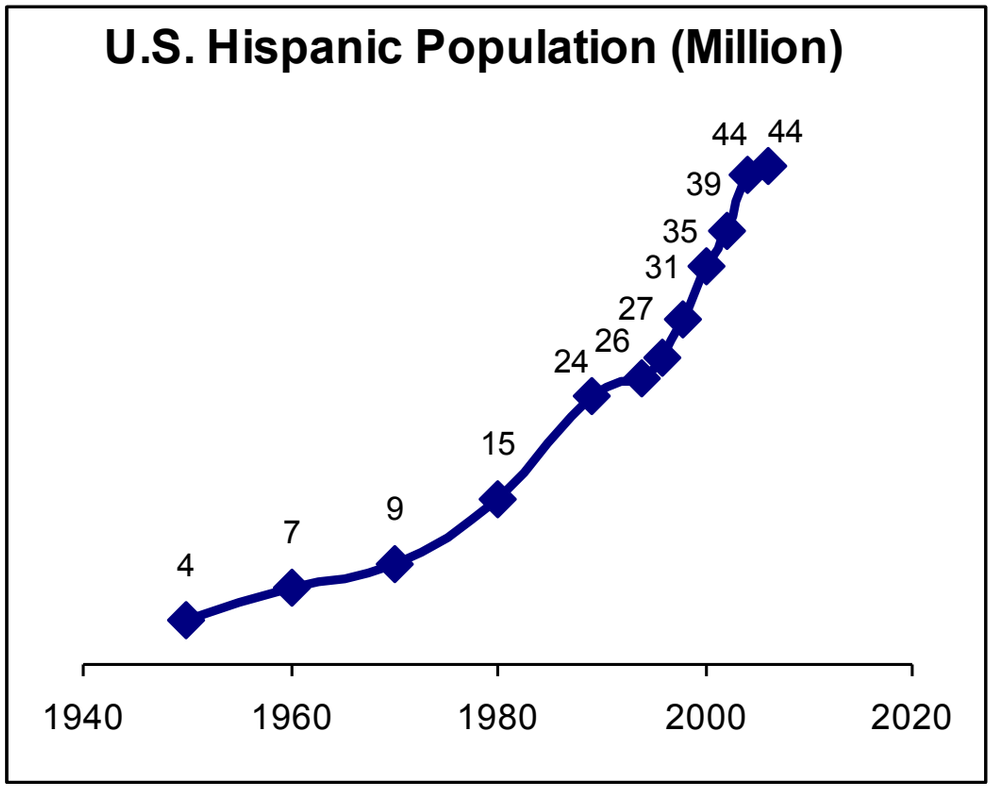
1. What is the US Hispanic Households segmentation problem? Why is it important to study the US Hispanic Households characteristics?

2. How to tackle the problem? - Previous works on clustering algorithms

3. Objectives

4. Project Timeline

The US Hispanic households characteristics (1/2)



US Hispanic population since 1950. Taken from the US Census Bureau

- Fast growing population
- Largest minority of the US
- <u>According with the US Census Bureau in 2050 hispanics will be one third of the US population</u>

The US Hispanic households characteristics (2/2)

- Different levels of acculturation
- Cultural heritage coming from more than 20 countries
- Different levels of literacy
- Different levels of affluence

Problem Statement (1/3)

The idea before the segmentation of the US Hispanic households is to find a set of subgroups of the population, so that that the households within each subgroup are homogenous enough to find common characteristics or patterns.

The previous problem can be thought as clustering problem as follows: let H be the set containing the Hispanic household sample, we want to find a set of subsets of H such that:

$$\cup_{i=1}^{k} C_i = H \quad (1)$$

$$\forall i = 1, \ldots, k \;\; C_i \neq \emptyset \quad (2)$$

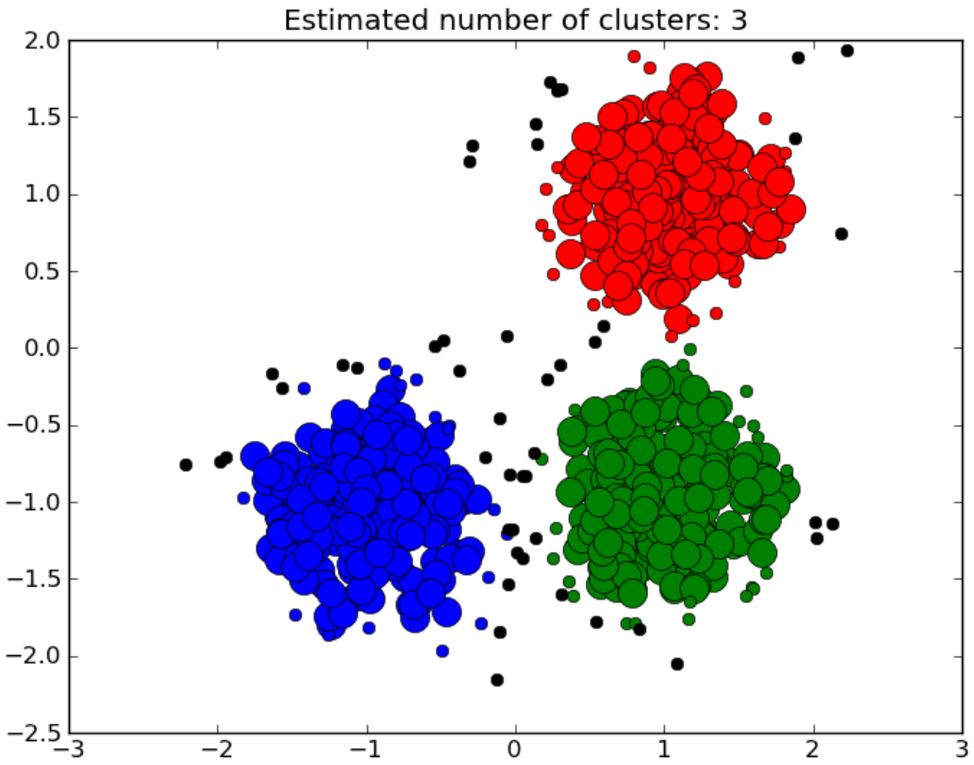
$$\forall i \neq j, \text{for } i, j = 1, \ldots, k \; C_i \cap C_j = \emptyset \quad (3)$$

Problem Statement (2/3)

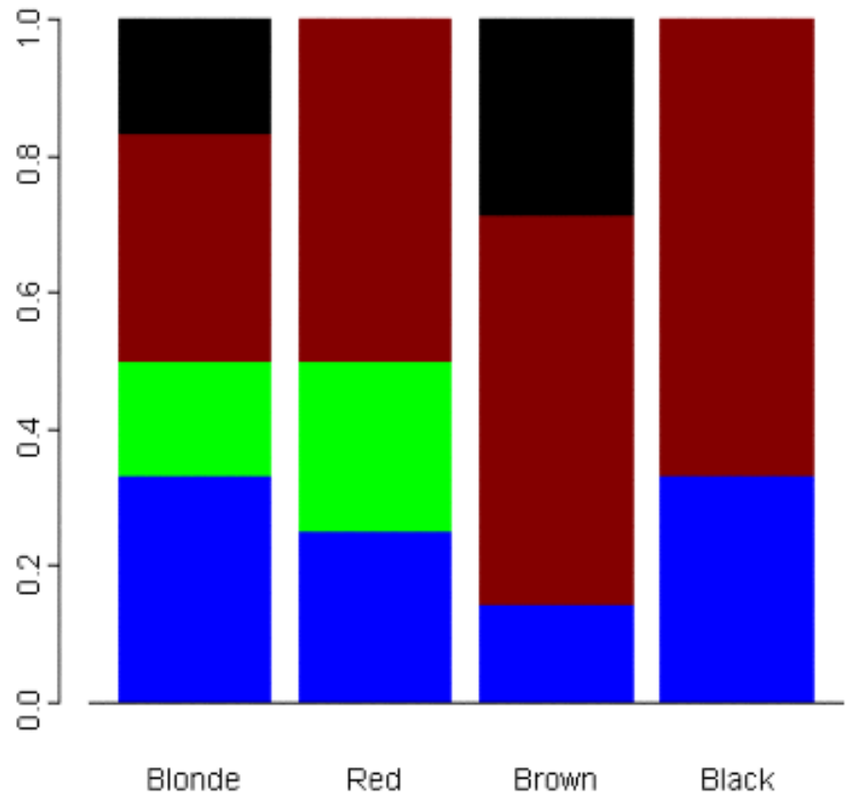
In addition, the set of clusters need to satisfy certain statistical properties to ensure homogeneity within the households of a given cluster and heterogeneity between the different clusters.

In order to generate the clusters fulfilling the previous conditions it becomes necessary to develop a clustering algorithm capable of generating valid solutions. It is also important to notice that the algorithm should take into account the the type of data describing the US Hispanic households.

Problem Statement (3/3)

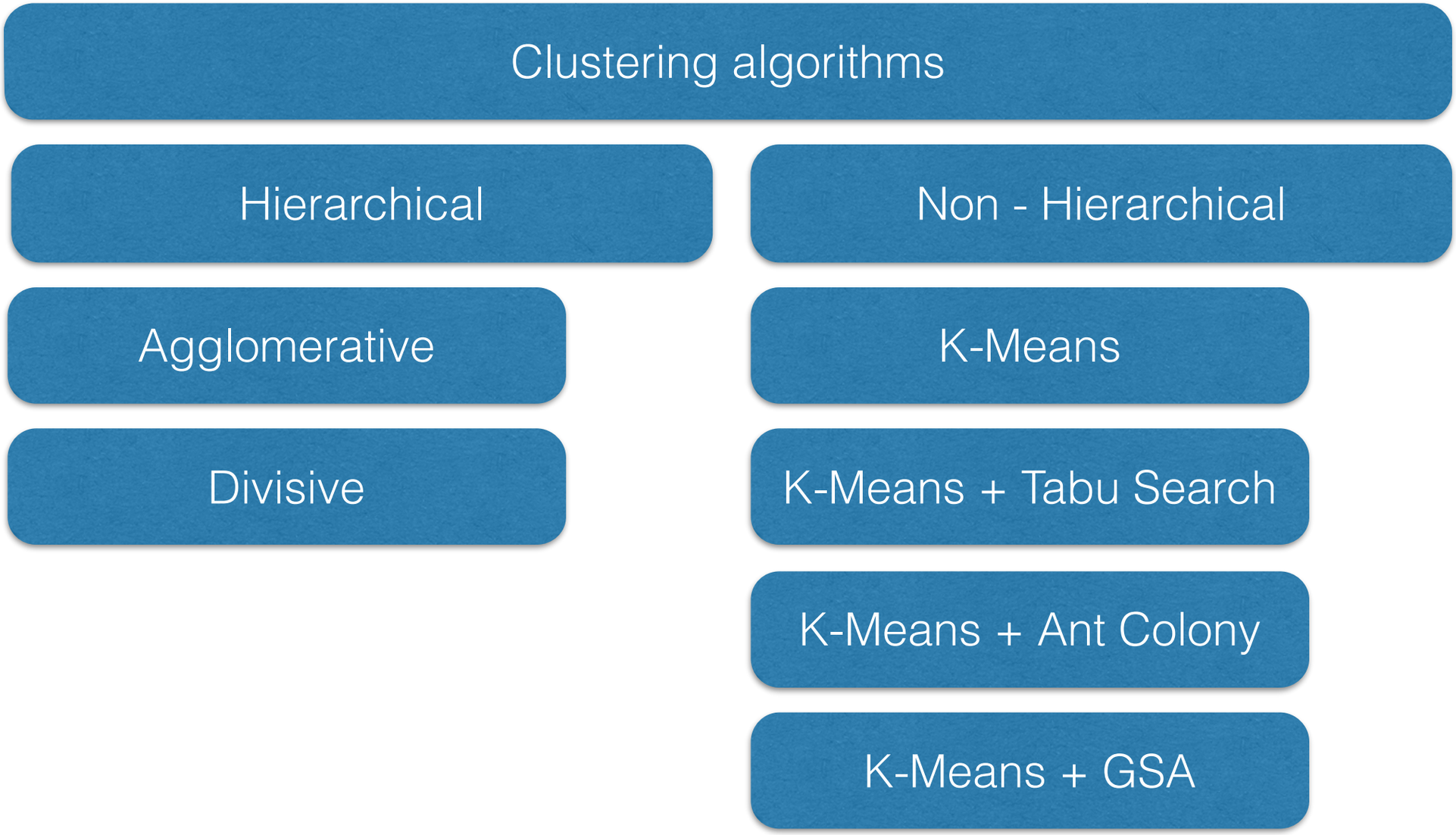


Numeric data



Categorical data

Previous works on clustering algorithms



Objectives

General Objective

Generate a useful classification of the Hispanic households in the U.S. to understand the Hispanic Household composition and its characteristics to support further marketing strategies.

Specific Objectives

1. Make a review of literature about the clustering algorithms

2. Prepare the Hispanic household data for the clustering analysis

3. Design and implement an adequate clustering algorithm for the Hispanic household data

4. Verify and validate the desired properties of the implemented algorithm (i.e. convergence)

5. Classify the Hispanic household data using the implemented algorithm and analyze the results.

Project Timeline

Activity	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15	Week 16
Data preparation and preprocessing														
Design and Implementation of the clustering algorithm														
Verification and validation of the algorithm and its properties														
Execution of the clustering analysis for the US Hispanic households														
Report write up and results discussion														

Thank you very much!

A CLUSTERING APPROACH FOR US HISPANIC HOUSEHOLDS SEGMENTATION

Juan Sebastián Marín-Delgado
jmarind@eafit.edu.co
Tutor: Francisco Zuluaga-Díaz
fzuluag2@eafit.edu.co