**Clustering analysis of mobility data**

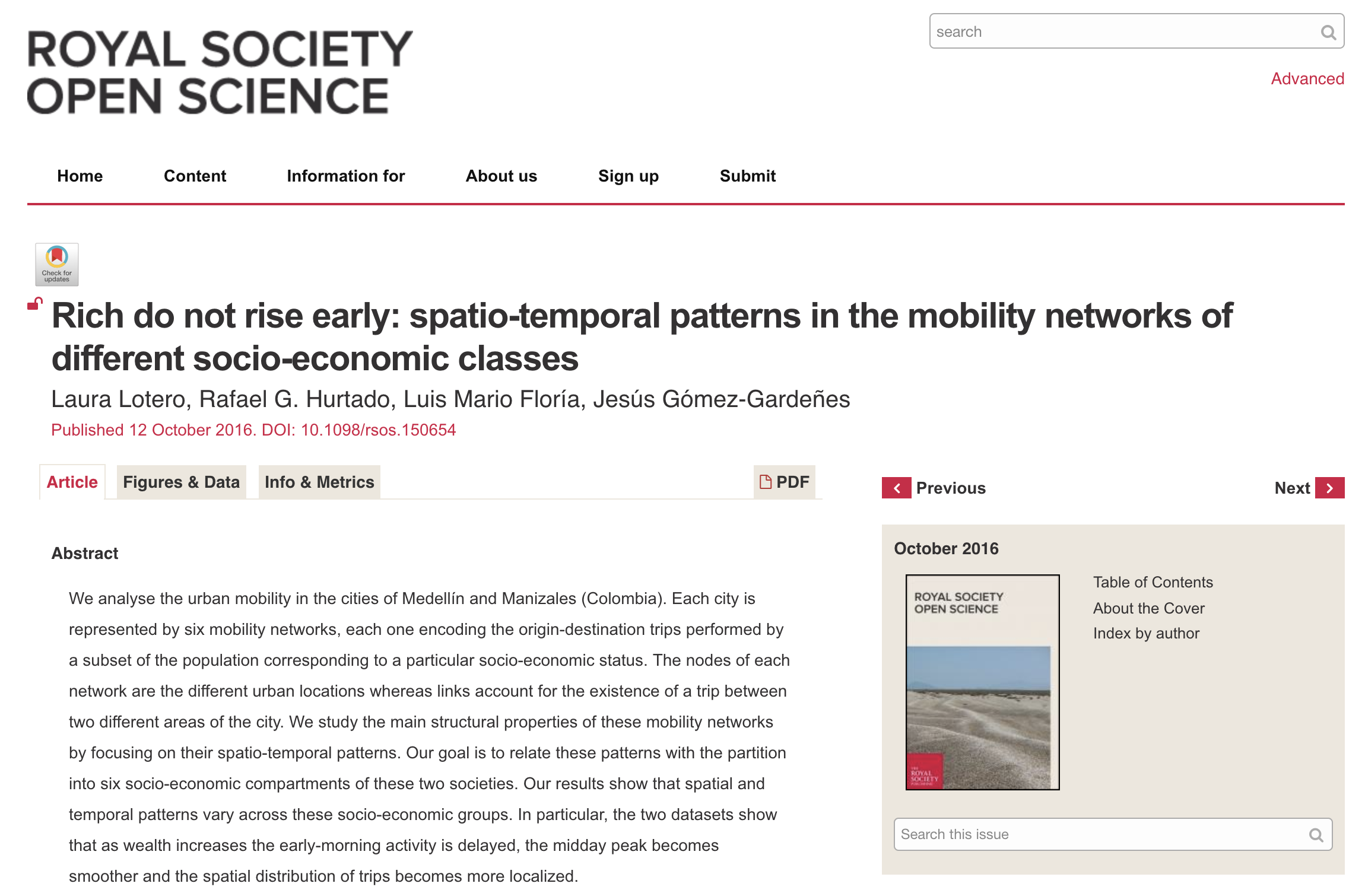


Figure from “Rich do not rise early: spatio-temporal patterns in the

mobility networks of different socio-economic classes ®

**Introduction to the work:**

“Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes” the authors study



**Objectives:**

- The aim of the work is to study whether the mobility information of individuals is enough to characterize (predict?) their social stratum, or whether different groups (within strata sot joining strata could appear), and discuss the implications of the findings.

- Additionally, we will perform analysis of importance of variables, to understand which variables are the most relevant in predicting the current (and also our own) social groups.

**The Data:**

We are going to focus just in the city of Medellín. Each data point aggregates information of the travel of a person.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Data type | Values | Explanation |
| Origin | Categorical | [1, 413] | Medellín areas |
| Destination | Categorical | [1, 413] | Medellín areas |
| Reason | Categorical | [1, 7] | Work, errands, … |
| Mean of transportation | Categorical | [1, 7] | Mean of transportation |
| Average time | Categorical | {AM, MD, PM, MN} | Average between origin and departure times |
| Time interval | Categorical | {AM, MD, PM, MN} | Interval of the day |
| Duration | Numerical |  | Duration of the travel in minutes |
| Distance | Numerical |  | Distance between origin and destination centroids of the areas |
| Age | Numerical |  | Age |
| Gender | Categorical | [1,2] | Gender |
| Strata\* | Categorical | [1,6] | Social strata |

\* We are not using this, we want to compare our clustering results to these categories.

The data set has 124,979 rows of individual travels.

The data is not balanced between stratas.

**How to:**

Unsupervised clustering for mixed data

Firstly we will try

<https://www.r-bloggers.com/clustering-mixed-data-types-in-r/>

* Create a measure of dissimilarity of data points using Gower similarity coefficient
* Cluster data using PAM (Partition around methoids) or Hiercarchical Clustering
* For PAM, we need to define the number of clusters (we will try validation metrics as Silhouette)

Other algorithms may be explored. Research on clustering of mixed data need to be carried out.

k-modes <https://github.com/nicodv/kmodes>

<https://arxiv.org/ftp/cs/papers/0603/0603120.pdf>

  k-prototypes (also here https://github.com/nicodv/kmodes)