Social Specialization of Space: Clustering Households on the French Riviera

Denis Pallez
Univ. Nice Sophia Antipolis,
I3S, UMR 7271, 06900 Sophia Antipolis, France
denis.pallez@unice.fr

Mathieu Serrurier
University Paul Sabatier
IRIT, UMR 5505
Toulouse, France
serrurie@irit.fr

Célia da Costa Pereira
Univ. Nice Sophia Antipolis,
I3S, UMR 7271, 06900 Sophia Antipolis, France
celia.pereira@unice.fr

Giovanni Fusco
Univ. Nice Sophia Antipolis,
Espace, UMR 7300, 06204
Nice, France
Giovanni.Fusco@unice.fr

Cristina Cao
Politecnico di Milano
DAStU
Milano, Italy
cristina.cao@mail.polimi.it

ABSTRACT

The aim of this paper is to estimate the extent of social specialization of residential space within the French Riviera metropolitan area. Unlike classical approaches, where social groups are pre-defined through given characteristics of households, our approach determines clusters of households inductively. Socio-demographic characteristics of households are thus measured through 16 different indicators. Clustering is then carried out through the optimization of two distinct criteria. Simulated annealing, simple and multi-objective Genetic Algorithm were adapted for this purpose and has produced pertinent results.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data Mining, Spatial databases and GIS*; I.5.3 [**Pattern Recognition**]: Clustering; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

Keywords

NSGA-II, Simulated Annealing, Metropolitan Areas

1. INTRODUCTION

Social specialization of residential space within an urban area is the concentration of households according to some characteristics like social status, demographic characteristics, ethnicity, etc. in different urban subspaces. Understanding the logics of social specialization is of course crucial in order to define policies like urban planning or housing.

Geographers have traditionally used two approaches in order to study social specialization of urban space. The first is to pre-select social groups defined along lines of social status, ethnicity and position in the life cycle. Spatial concentration of these groups is later analyzed in order to detect eventual segregation patterns within the city. The second approach aims at characterizing the content of the resident population of neighborhoods without targeting a specific social group but by taking into account the whole population and all descriptors of households at the same time. Unlike these classical approaches, we want to both determine spatial concentration of specific groups of population in the urban space and to take into account all descriptors of households at the same time. This poses the problem of clustering households according to two conflicting criteria: maximum homogeneity of human content and maximum spatial difference within the metropolitan area. Classical clustering can not be directly applied due to optimization of two conflicting criteria. This is the reason why we adapt an evolutionary and a simulated annealing algorithm and apply them to a data set of households of the French Riviera extracted from the CERTU Household Mobility Survey of 2008.

GECCO '15 July 11-15, 2015, Madrid, Spain

ACM ISBN 978-1-4503-3488-4/15/07.
DOI: http://dx.doi.org/10.1145/2739482.2764665

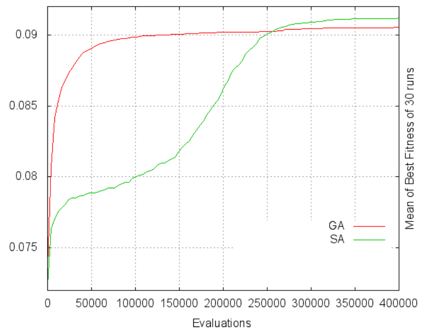
2. OUR FRAMEWORK

Given a set of n households $I = \{x_1, \dots, x_j, \dots, x_n\}$ where $x_j = \{x_{j1}, \dots, x_{j16}\}$ represents the j-th household, and each x_{ji} represents its ith feature value. We name w_i the weight associated with the i-th attribute and $coem_j$ the weight corresponding to the j-th household in the sample.

In this paper we consider the clustering based on medoids: each cluster $C_{i \in 1..k}$ is defined by a center $m_i \in I$. One common measure used for clustering is the intra-cluster distance between medoids and items and can be defined as:

$$Intra(C) = \frac{\sum_{i=1}^{k} \sum_{x_j \in C_i} coem_j \times d(m_i, x_j)}{\#I}, \qquad (1)$$

where $d \in [0, 1]$ represents the distance between two items weighted by w_i. Moreover, $\#I = \sum_{j=1}^{n} coem_j$. In the following the cardinality of a subset $A \subset I$ will be the sum of the *coem*s of the items it contains, i.e. $\#A = \sum_{x_j \in A} coem_j$. However, the minimization of equation 1 is not sufficient to determine clusters of households that are spatially opposed within the metropolitan area. In order to ensure that socio-demographic differences among clusters also correspond to



(a) Best single objective fitness of GA and SA on 30 runs

MOGA PF_17				WDI =	0.231
Social Status	SINGLE ADULTS	COUPLES (with children)	COUPLES WITH CHILDREN	SINGLE RETIREES	COUPLES OF RETIREES
high	G5 - 0.23 ; 7.10%		G1 - 0.25 ; 19.0%		G2 - 0.27 ; 6.25%
medium high		G4 - 0.20 ; 5.67%		G3 - 0.22 ; 13.7%	
medium low			G0 - 0.24 ; 9.03%	G9 - 0.21 ; 10.7%	G7 - 0.16 ; 13.1%
low	G8 - 0.22 ; 10.9%				G6 - 0.33 ; 4.34%
particular profiles					

MOGA PF_12				WDI =	0.237
Social Status	SINGLE ADULTS	COUPLES (with children)	COUPLES WITH CHILDREN	SINGLE RETIREES	COUPLES OF RETIREES
high	G0 - 0.24 ; 6.73%		G6 - 0.25 ; 20.3%		G7 - 0.24 ; 8.99%
medium high				G5 - 0.22 ; 21.1%	
medium low			G3 - 0.20 ; 9.65%	G1 - 0.26 ; 4.76%	G8 - 0.19 ; 8.58%
low	G9 - 0.22 ; 10.0%				G2 - 0.32 ; 4.68%
particular profiles			G4 - 0.29 ; 5.02%		

MOGA PF_48				WDI =	0.246
Social Status	SINGLE ADULTS	COUPLES (with children)	COUPLES WITH CHILDREN	SINGLE RETIREES	COUPLES OF RETIREES
high	G4* - 0.32 ; 3.16%		G7 - 0.24 ; 21.8%		G1 - 0.22 ; 12.9%
medium high	G9 - 0.23 ; 7.43%			G2 - 0.24 ; 17.1%	
medium low	G0 - 0.20 ; 9.21%		G5 - 0.27 ; 8.53%	G3 - 0.28 ; 5.09%	G6 - 0.18 ; 11.8%
low					
particular profiles				G8 - 0.54 ; 2.88%	

(b) Geographic interpretation of clusters produced by NSGA-II

Figure 1: Results

spatial difference in sectors of residence, we create a new criteria based on dissimilarity index by Duncan and Duncan [2]. It indicates the percentage of a given target population that should move in order to spread it evenly within the metropolitan space. Our Weighted Dissimilarity Index is a weighted average of Duncan and Duncan index over the different clusters identified within the metropolitan population and defined as $WDI(S, C) =$

$$\sum_{j=1}^{k} \frac{\#C_j}{2\#I} \times \sum_{i=1}^{q} \left| \frac{\#(S_i \cap C_j)}{\#C_j} - \frac{\#S_i - \#(S_i \cap C_j)}{\#I - \#C_j} \right|, \quad (2)$$

where $S = \{S_1, \ldots, S_q\}$ be the q-partition of I in sectors. Household clustering will thus be carried out by simultaneously minimizing Equation 1 and maximizing Equation 2.

3. EXPERIMENTS

Simulated Annealing (SA) [4] and Genetic Algorithm (GA) have been adapted for optimizing a single objective combining both criteria using $f(S, C) = \frac{WDI(S,C) - Intra(C)}{2}$. For SA, a state is a vector of k distinct elements of I that correspond to the medoids of the cluster. The neighborhood of a state $s = \{c_1, \ldots, c_k\}$ is obtained by switching one of the medoids in s with an item in I. In order to be comparable with other strategies, we perform 400000 evaluation by choosing $NBLOOP = 2000$ and $NBSTEP = 200$. We fix $t_0 = 0.1$ and $r = 0.995$. s_0 is chosen randomly. Results are depicted in figure 1(a). For GA, a medoid-based representation clearly explained in [3] is used with reset mutation, tournament selection of size 10, a population size of 200 and a one point crossover. We then adapted NSGA-II [1] where objectives are considered independently.

4. DISCUSSIONS

Socio-demographic content of clusters extracted from Pareto front obtained by NSGA-II (PF17, PF12,PF48) is analyzed and presented in figure 1(b). Colors represent both household configuration and social status of the average cluster profiles in every solution. The combination of the five configurations with the four social strata produces twenty possible cases. A few additional cases account for particular profiles. Cluster labels G0-G9 are followed by the Dissimilarity Index (their weighted average gives the WDI value) and the cluster share within the total population. Intermediate Pareto front solutions are robust as WDI grows. For every configuration of households, two different social positions are normally identified. However, only PF17 identifies a cluster of couples (sometimes with children) of medium-high social status. Particular profiles with higher DI are detected by PF12 and PF48: couples of retirees of low social status and very segregated single retirees in social housing, respectively. Extreme solutions on the Pareto front are less interesting and even less pertinent when WDI is maximized at the expense of cluster internal heterogeneity.

5. CONCLUSION AND FUTURE WORK

For understanding social specialization on the French Riviera, we have adapted an NSGA-II algorithm and proposed a framework for clustering with a double optimization scope: minimizing intra-cluster distances and maximizing Weighted Dissimilarity Index seems to us the only way to empirically determine which categories of population are most opposed from a spatial point of view. The produced results seem in good agreement with expert knowledge of the social structure of the French Riviera. For the geographer, the task is now to project clustering results in geographic space in order to see what the ID values of the different clusters mean in terms of social morphology of the metropolitan area.

6. ACKNOWLEDGMENT

This work is a result of "Géo-Incertitude" project (2014-15) funded by CNRS within the research program HumaIn.

7. REFERENCES

[1] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6(2):182–197, Apr 2002.

[2] Otis Dudley Duncan and Beverly Duncan. A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):pp. 210–217, 1955.

[3] E.R. Hruschka, R.J.G.B. Campello, A.A. Freitas, and A.C.P.L.F. de Carvalho. A survey of evolutionary algorithms for clustering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(2):133–155, March 2009.

[4] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.