Universitat de Girona
Escola Politècnica Superior

DECISION SUPPORT METHODS FOR GLOBAL OPTIMIZATION

by
Ferran Torrent Fontbona

Advisors
Beatriz López Ibáñez
Víctor Muñoz Solà

MIIACS
September 2012
Girona

SUMMARY

- **Introduction**
  - Motivation
  - Objectives
  - The data
- State of the art
- Clustering
- Optimization
- Conclusions
- Future work

Motivation

- Globalization of the sport events
- Several simultaneous sport events

→ Barman decision problem →

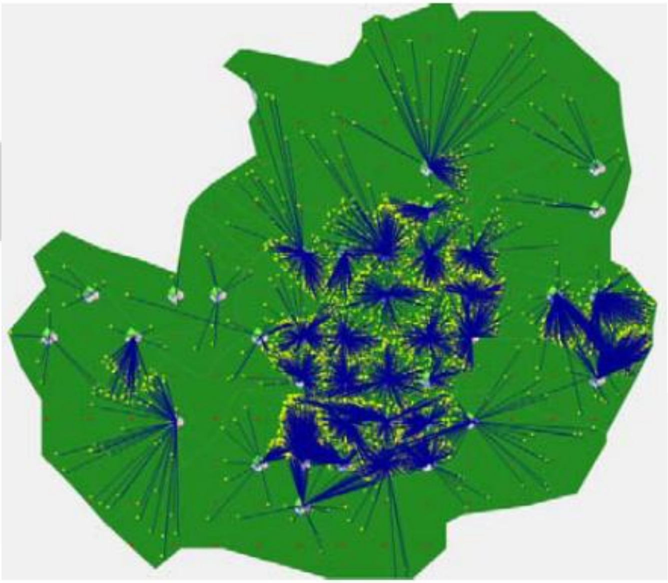
80 people wants match 1
20 people wants match 2

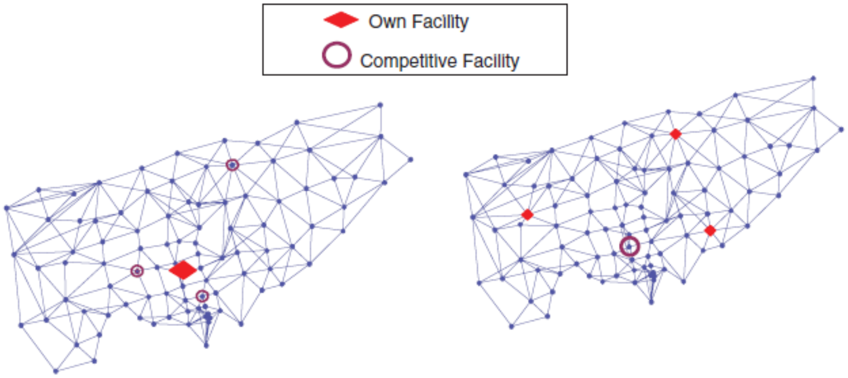
→ 10 bars broadcast match 1 → 8 people/bar

→ 8 bars broadcast match 1
2 bars broadcast match 2 → 10 people/bar

MOTIVATION. MATCHING PROBLEM

- Location-allocation
  - Determine optimal location for one or more facilities that will service demand for a given set of points
  - Every facility offers the same service
  - Customers positions are known
  - Complexity → $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ where $\begin{cases} n \rightarrow & number\ of\ possible\ positions \\ k \rightarrow & number\ of\ facilities \end{cases}$





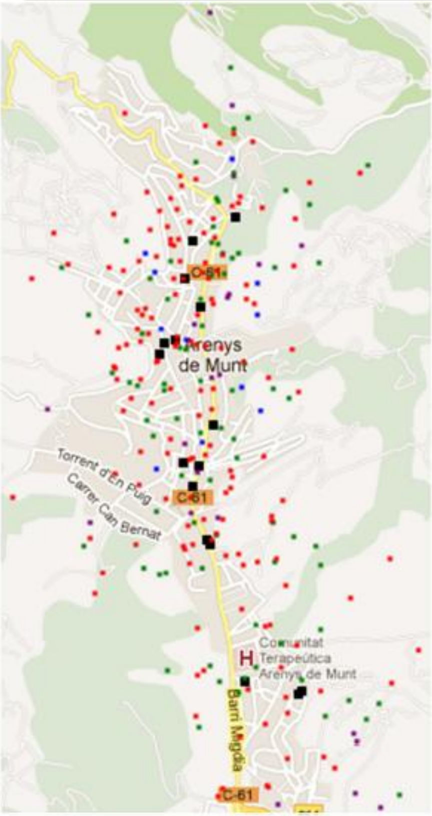
MOTIVATION. OUR PROBLEM

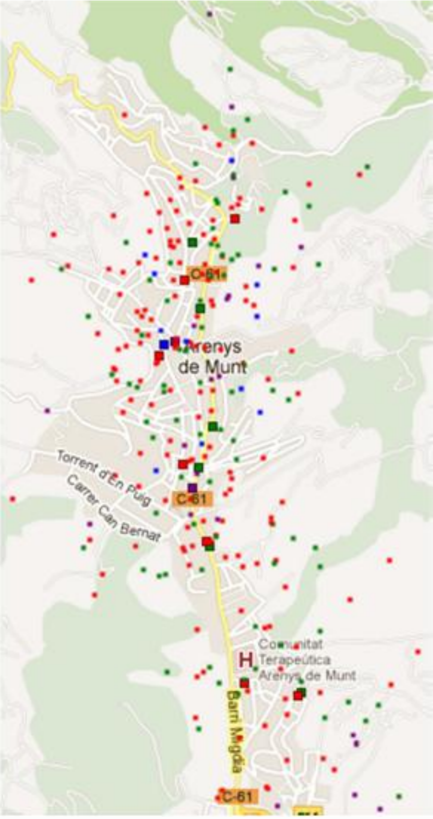
- Immobile location-allocation
  - Given a set of facilities with known positions and a demand with known positions, determine the optimal service each facility has to offer
  - Facilities (bars) cannot be moved and their positions are known
  - Each customer desire a single service (match) from a set and it is known
  - Customers' positions are known
  - Complexity → $(N_{matches})^{N_{bars}}$

- Problem dimensionality
  - Most research does not deal with problems of the same complexity/size (the system has to deal with bars from around the world)

Division of the problem into subproblems

$$k \cdot (N_{matches})^{N_{bars}/k}$$

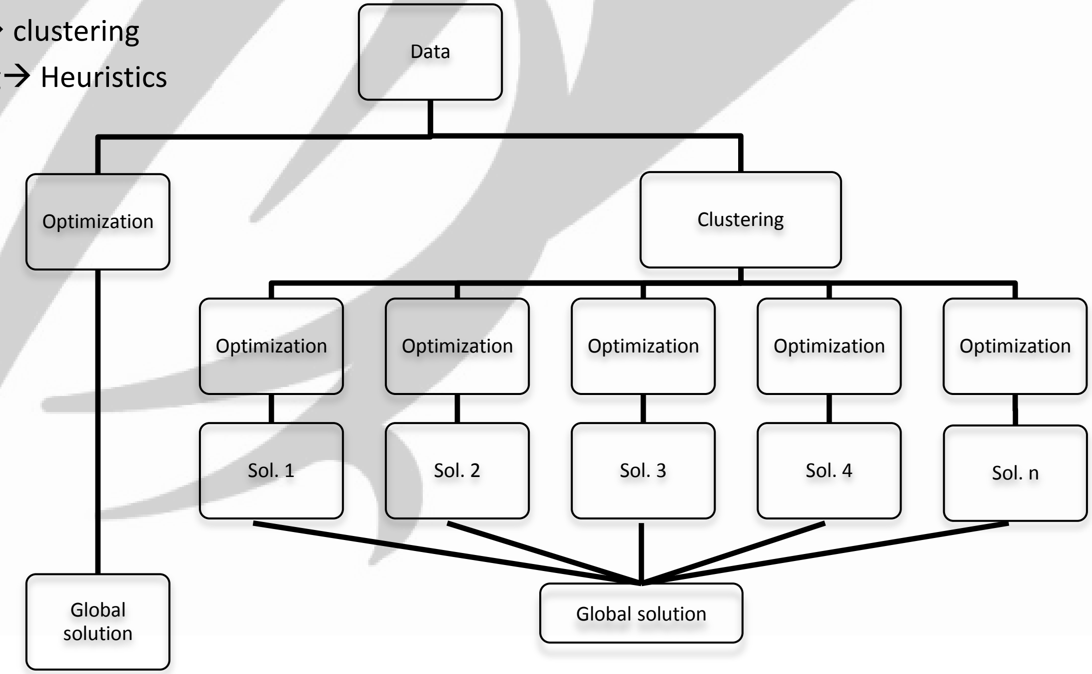




Objectives

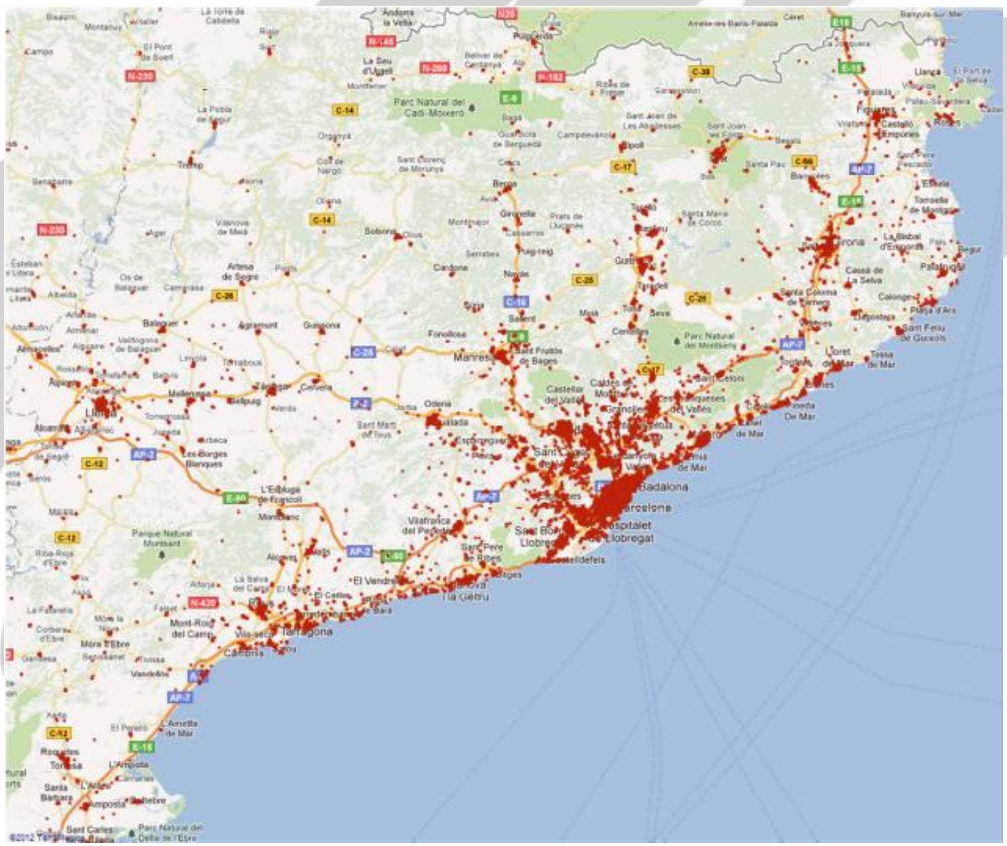
- Hypothesis
  - We can approximate the location-allocation solution regarding bars problem by dividing the dataset converting the initial problem into several of easier subproblems.
  - Assumption: geographical distance is a key of the problem and clustering divides the problem according the distance.

- Objectives
  - Divide the problem into sub-problems → clustering
  - Location-allocation (sub)problem solving→ Heuristics
  - Experimental tests



THE DATA

- 15578 bars from Catalunya taken from *Páginas Amarillas*
- Customers are randomly generated from a list of matches

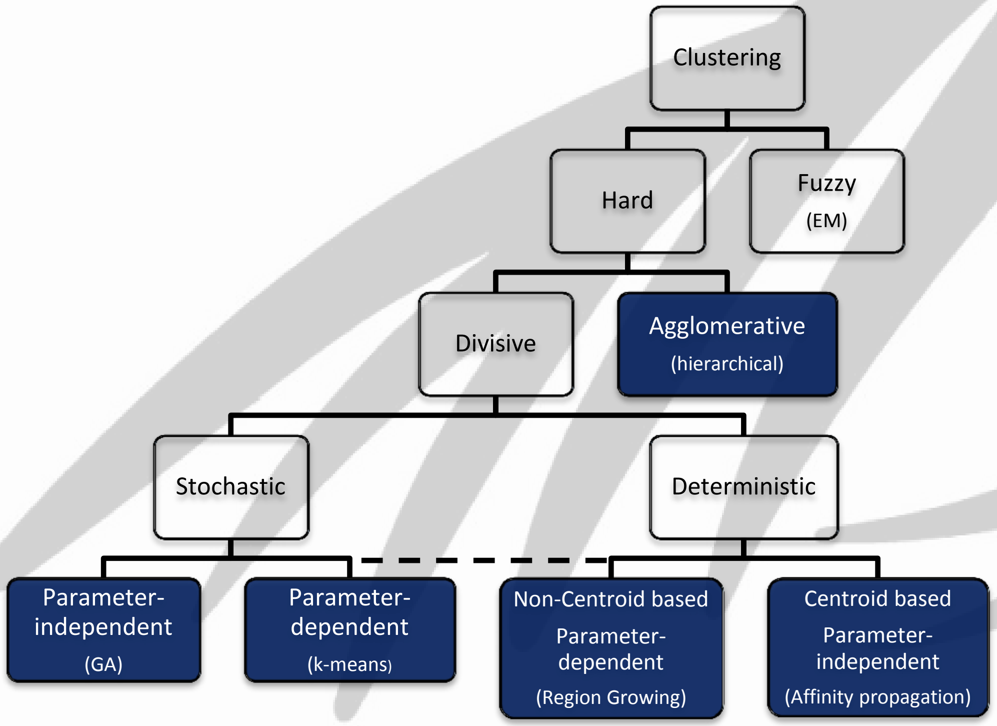


SUMMARY

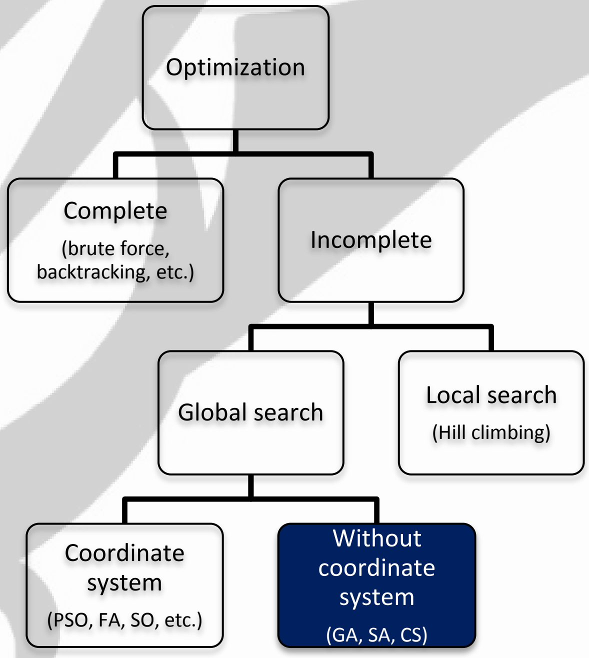
- ✓ Introduction
- ➢ **State of the art**
  - – Clustering
  - – Optimization
- Clustering
- Optimization
- Conclusions
- Future work

State of the Art

Clustering



Optimization

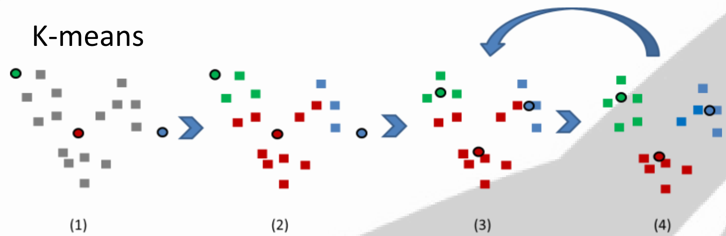


SUMMARY

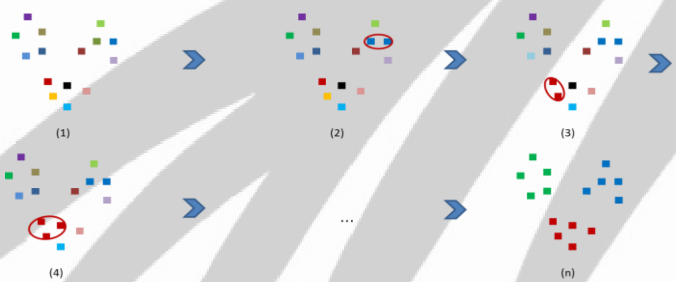
- ✓ Introduction
- ✓ State of the art
- ➢ **Clustering**
  - Algorithms
  - Results
- Optimization
- Conclusions
- Future work

CLUSTERING

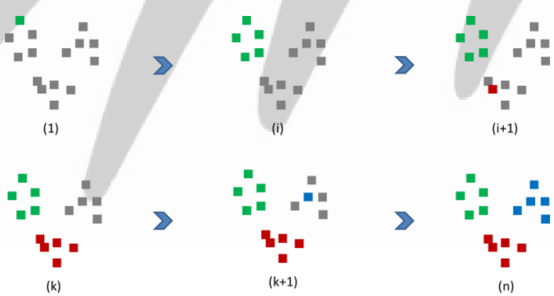
- Algorithms
  - K-means



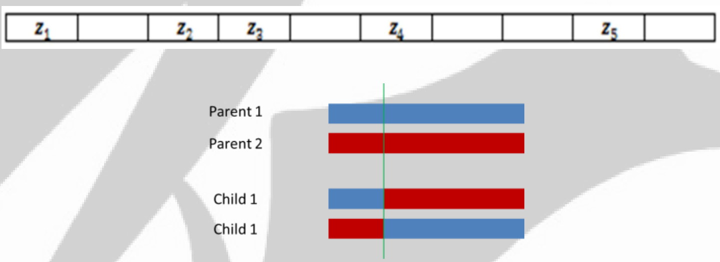
  - Hierarchical clustering



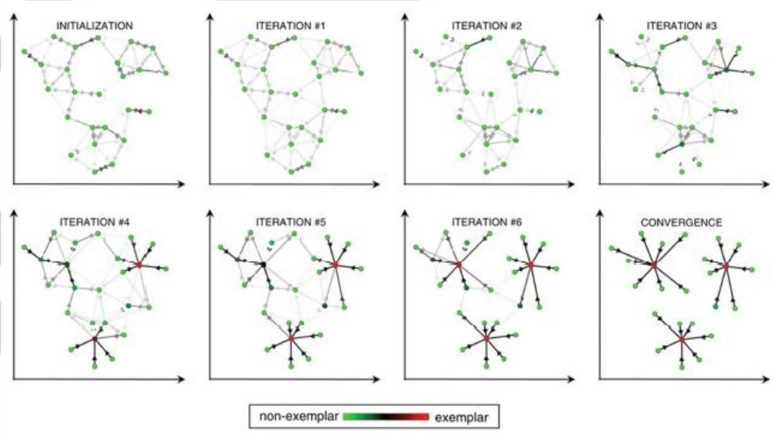
  - Region Growing



  - Genetic algorithms based clustering

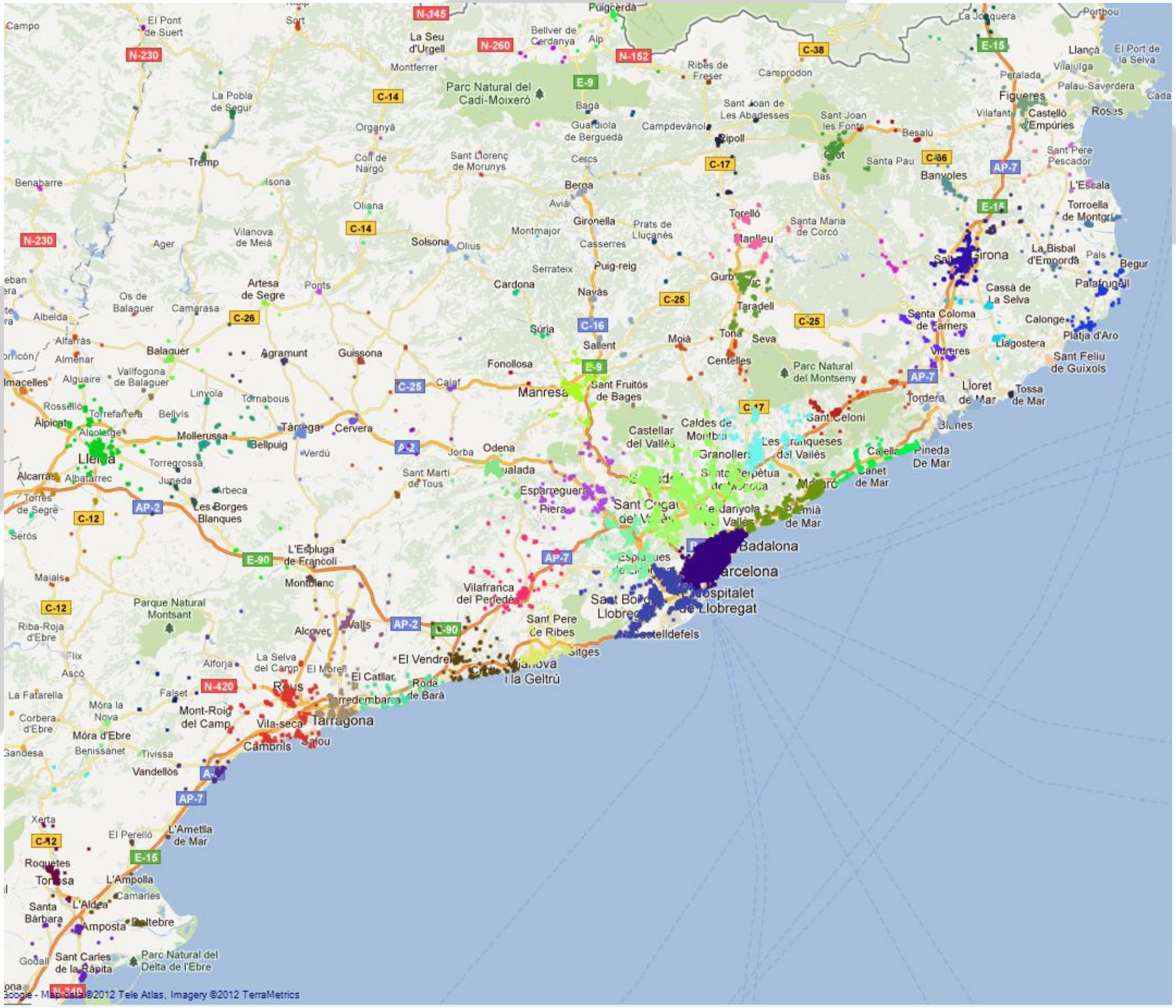


  - Affinity propagation



Clustering Results

- Hierarchical clustering



Clustering Results

- Initial complexity → $(N_{matches})^{N_{bars}} = 3^{15578} \cong 4 \cdot 10^{7432}$

Algorithm	Expended time (s)	Calinski Index	DB Index	Number of clusters	Number of minimal clusters	Smallest cluster size	Largest cluster size	Complexity
k-means (setting elements as initial centroids)	578	28955.66	0.717	896	27	1	**59**	$\mathbf{10^{31}}$
k-means (empty clusters resignation)	1170	**50166.93**	0.499	444	74	1	1001	10^{480}
Lloyd's algorithm	395	21958.88	0.698	17	1	137	3423	10^{1633}
Region growing $D_{max} =$ 1km	**6**	2614.59	0.228	1095	**521**	**1**	5885	10^{2810}
Region growing $D_{max} =$ 2km	12	1182.52	**0.224**	707	288	1	8202	10^{3916}
Region growing $D_{max} =$ 5km	37	**430.88**	0.383	280	93	1	**10733**	10^{5123}
Hierarchical clustering	**36636**	16592.55	0.472	139	10	1	4487	10^{2142}
Genetic clustering	4575	15911.56	**0.757**	14	1	366	2305	10^{1100}
Affinity propagation	3892	27037.92	0.665	92	**1**	**18**	690	10^{331}
	[↓,↑]	[↓,↑]	[↓,↑]			[↓,↑]	[↓,↑]	[↓,↑]

SUMMARY

- ✓ Introduction
- ✓ State of the art
- ✓ Clustering
- ➢ **Optimization**
  - Mathematical model
  - Genetic algorithms
  - Simulated annealing & cuckoo search
  - Results
- Conclusions
- Future work

LOCATION-ALLOCATION

- Mathematical model

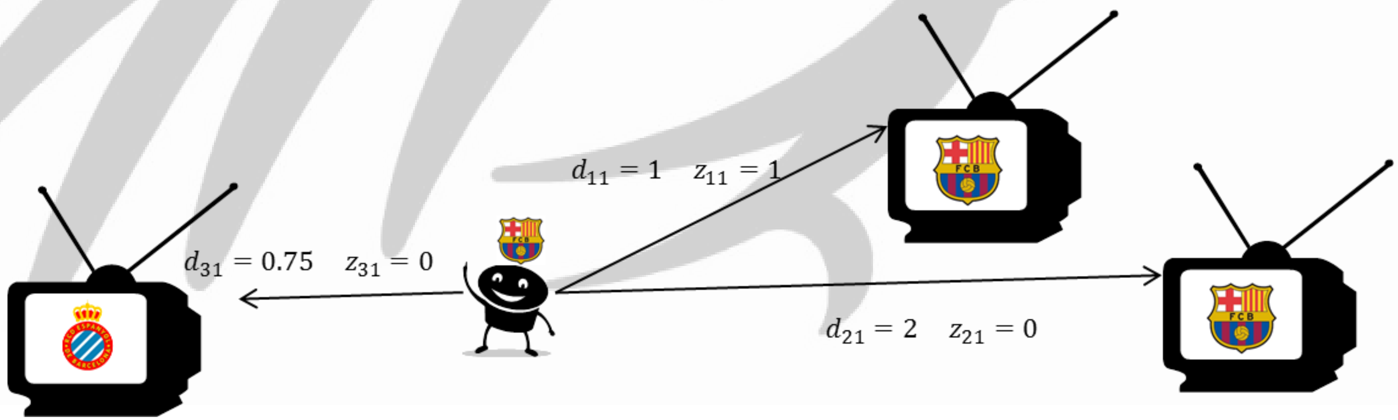
$$\max_{z_{ij}^q} \left\{ \sum_{i=1}^{N_{bars}} \sum_{j=1}^{N_{customers}} \frac{z_{ij}^q}{1+d_{ij}^2} \right\}$$

Subject to

$$\forall_i \quad \sum_{j=1}^{N_{customers}} z_{ij}^q \le C_i$$

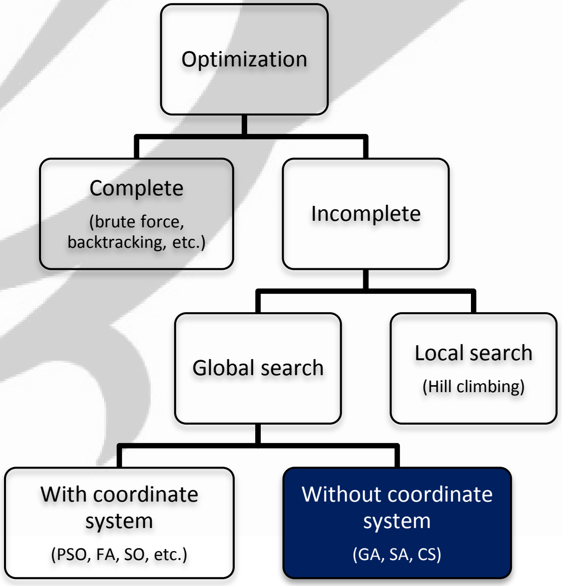
$$\forall_j \quad \sum_{i=1}^{N_{bars}} z_{ij}^q \le 1$$

$$x_i^q \neq M_j \rightarrow z_{ij}^q = 0, \qquad x_i^q, M_j \in [1, \cdots, N_{matches}]$$



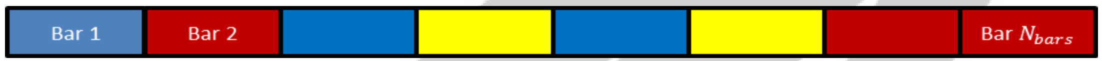
Optimization Methods

- × Complete methods → the number of solutions to be explored is too big
  - Brute force, depth-first search, breath-first search, backtracking, etc.
- × Local search methods → many local optimums
  - Gradient based methods, hill climbing
- × Heuristics with coordinate systems → non-coordinate solution space!!
  - PSO, FA, SO, etc.
- ✓ Heuristics with non-coordinate systems → find good solutions in a limited amount of time
  - GA, SA, CS



Genetic Algorithms

- Chromosome



- Mutation
  - Probability μ_m to change the match
- Crossover
  - Single point crossover
- Fitness

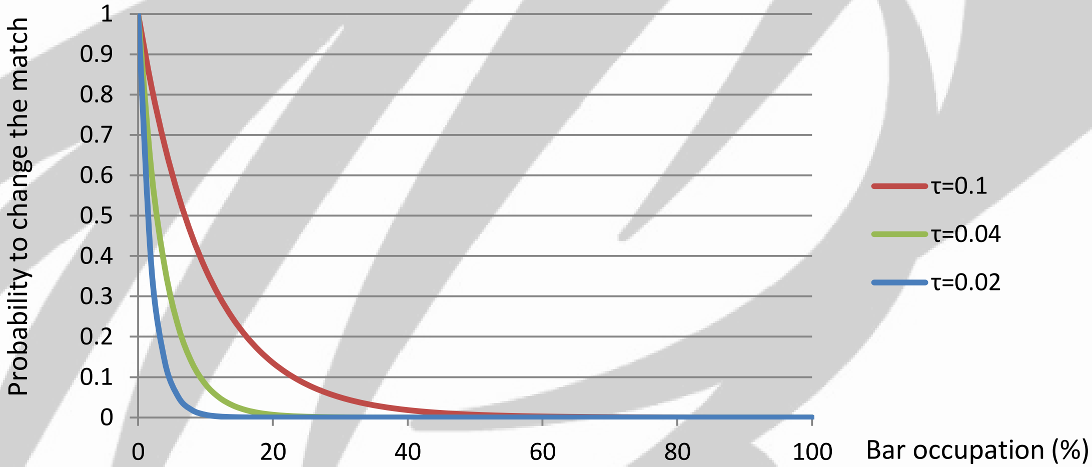
$$Fitness(q) = \sum_{i=1}^{N_{bars}} \sum_{j=1}^{N_{customers}} \frac{z_{ij}^{q}}{1 + d_{ij}^{2}}$$

- Selection
  - Roulette rule

SIMULATED ANNEALING & CUCKOO SEARCH

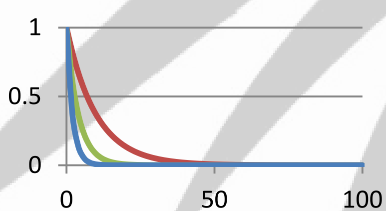
- Non-coordinate search space → **Need of a new neighborhood function**
  - Each bar have different chances to change its match depending on the expected number of customers → Exponential probability function
  - Different exponential function depending on the features of the problem

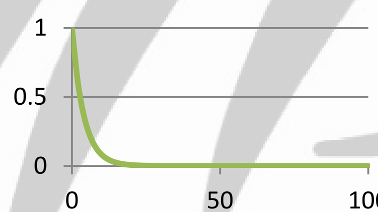
$$P\left(\text{change the match of the ith bar}\right) = e^{-\frac{o_i}{\tau}}$$

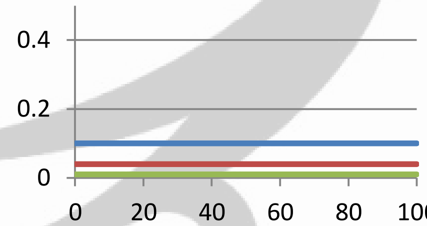


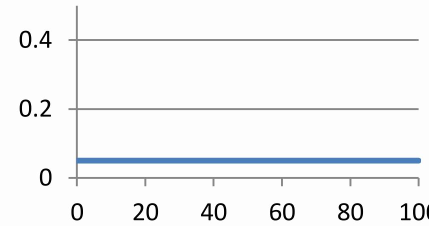
SIMULATED ANNEALING & CUCKOO SEARCH

Exponential probability with variable τ			Exponential probability with $\tau = 0.05$			Variable uniform probability			Constant uniform probability		
E	% of allocated customers	% of bars with occupation < 4%	E	% of allocated customers	% of bars with occupation < 4%	E	% of allocated customers	% of bars with occupation < 4%	E	% of allocated customers	% of bars with occupation < 4%
217.04	95.33	0	211.34	94.00	0	214.45	95.00	0	216.15	93.00	0
104.43	97.82	1	103.85	98.55	3	103.04	98.55	2	104.01	96.38	3
1223.49	99.43	0	1218.94	98.93	0	1221.93	98.93	0	1218.18	98.93	2
616.49	99.86	3	616.55	100	3	614.95	99.86	5	613.67	99.86	6
2010.62	100	0	2013.74	100	1	2005.71	100	8	2007.23	100	13
996.03	100	12	994.11	100	11	993.98	100	19	991.81	100	23
5579.03	99.83	1	5571.28	99.71	3	5535.93	99.73	48	5531.09	99.68	41
2622.78	99.86	20	2622.36	99.89	28	2612.07	99.96	89	2606.94	99.75	91









Location-Allocation Results

- SA achieves the best solutions
- Individual LA is the fastest method but also finds the worst solutions
- SA and CS spend the same amount of time approx.

Number of facilities	Fitness				% of allocated customers				% of facil. with occupation < 4%				Elapsed time (s)			
	Individ.	GA	SA	CS	Individ.	GA	SA	CS	Individ.	GA	SA	CS	Individ.	GA	SA	CS
8	**81.39**	**109.56**	108.27	107.30	**56.73**	**79.30**	78.13	78.13	**0.00**	4.29	**0.00**	**0.00**	**0.000**	**0.467**	0.129	0.136
18	**170.38**	279.91	**281.86**	278.35	**51.39**	94.16	95.72	**95.82**	**0.00**	1.11	**0.00**	1.11	**0.001**	**3.103**	0.662	0.702
42	**438.26**	707.69	**723.27**	713.74	**56.94**	**99.88**	99.83	99.55	**0.00**	12.61	**0.00**	0.48	**0.009**	**17.164**	4.140	4.083
46	**427.11**	681.92	**706.08**	696.38	**55.50**	98.17	**99.68**	99.54	**2.17**	13.06	**2.61**	3.48	**0.009**	**11.741**	2.440	2.878
48	**479.4**	824.50	**838.18**	832.65	**53.85**	99.50	**99.58**	**99.58**	**0.00**	4.58	**0.00**	**0.00**	**0.011**	**22.155**	5.660	6.146
50	**484.39**	754.45	**776.96**	768.91	**57.10**	97.58	**97.97**	97.94	**2.00**	12.40	**0.00**	1.20	**0.004**	**16.409**	4.067	4.323
72	**622.92**	1057.11	**1079.42**	1074.73	**54.89**	98.89	**98.97**	98.89	**0.00**	4.58	**3.06**	3.06	**0.021**	**34.486**	11.088	11.553
127	**1389.85**	2374.754	**2421.44**	2404.44	**55.58**	**100.00**	**100.00**	**100.00**	**0.79**	14.80	**0.16**	1.57	**0.028**	**159.720**	50.617	48.039
313	**3019.05**	5144.42	**5258.10**	5238.18	**55.75**	99.58	**99.75**	99.74	**0.32**	21.15	**0.58**	3.07	**0.136**	**712.152**	293.865	288.316
1495	14660.55	-	**25826.85**	25762.79	55.91	-	99.97	**99.99**	**0.07**	-	0.54	3.28	**3.571**	-	5285.298	4934.568

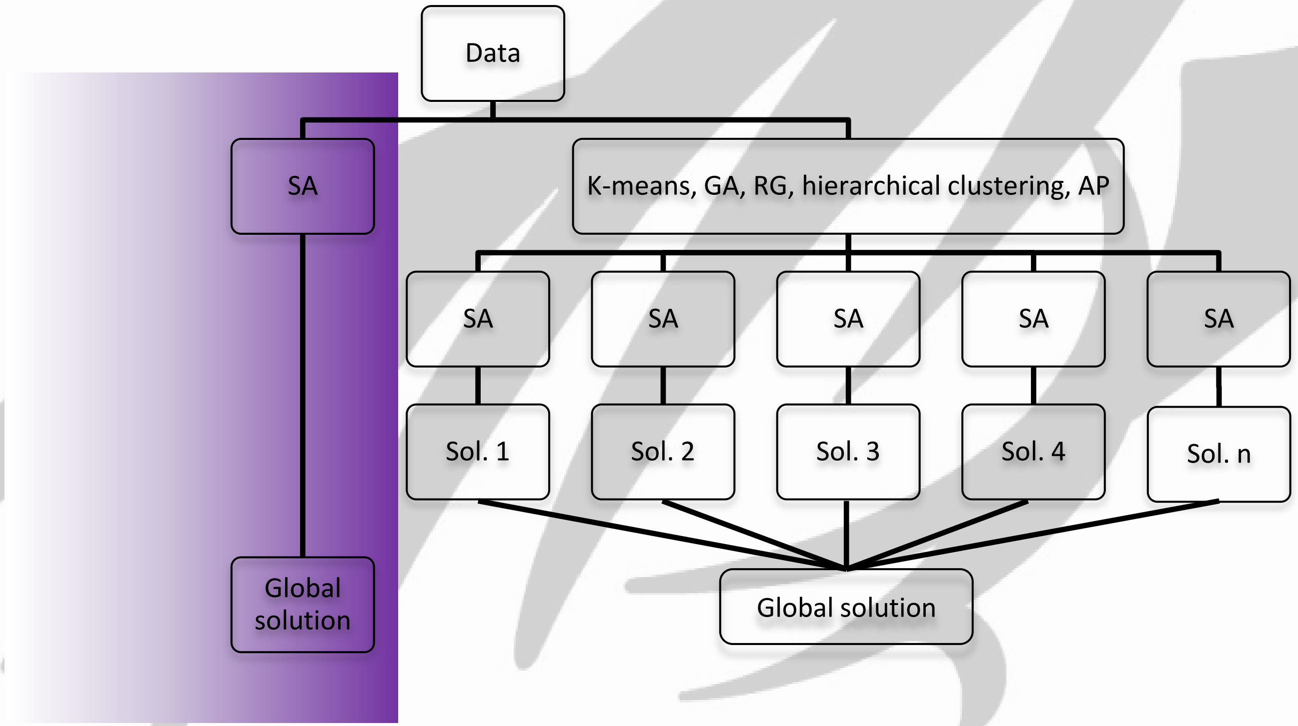
Clusters from k-means clustering

LOCATION-ALLOCATION RESULTS

- What if we initialize SA with the solution found by individual LA?

Number of facilities	Fitness		% of allocated customers		% of facilities. with occupation < 4%		Elapsed time (s)	
	SA	Individual LA & SA	SA	Individual LA & SA	SA	Individual LA & SA	SA	Individual LA & SA
18	257.42	**259.89**	97.38	**97.52**	**0.00**	**0.00**	**0.591**	0.612
42	704.87	**707.57**	**99.88**	**99.88**	**2.38**	**2.38**	**4.546**	4.579
72	1222.67	**1229.38**	98.92	**98.95**	**1.39**	**1.39**	14.747	**14.549**
127	2234.65	**2242.62**	**100.00**	**100.00**	2.76	**0.79**	49.893	**49.237**
313	5068.38	**5077.28**	99.77	**99.83**	2.08	**1.28**	**299.293**	299.298
1495	26229.77	**26259.56**	**99.99**	99.97	1.27	**0.54**	**5976.626**	6079.012

LOCATION-ALLOCATION RESULTS



LOCATION-ALLOCATION RESULTS

Technique	Dataset 1 (459 bars)			Dataset 2 (1925)		
	Num. (max) clust	Fitness	Time (s)	Num. (max) clust	Fitness	Time (s)
Non-clustered		7816.24	569.964		28778.06	2261.027
Genetic	8 (234)	7624.84	196.576	18 (548)	29041.34	669.030
Hierarchical	8 (234)	7632.91	205.237	48 (395)	29160.86	507.649
k-means (empty clusters resignation)	125 (55)	6311.87	16.456	185 (131)	28877.25	200.576
k-means (setting elements as initial centroids)	159 (53)	6120.70	14.702	834 (39)	23972.13	76.919
Lloyd's alg.	170 (44)	5983.05	25.618	654 (39)	25306.32	93.778
RG $D_{max} = 0.1$ km	172 (73)	5968.20	20.947	1082 (75)	22371.45	77.392
RG $D_{max} = 0.2$ km	71 (248)	7113.03	188.358	770 (264)	24888.39	205.940
RG $D_{max} = 0.5$ km	14 (405)	7726.98	512.064	401 (405)	28192.99	611.308
RG $D_{max} = 1.0$km	2 (457)	7801.64	569.461	258 (473)	29091.78	679.154
Affinity propagation	20 (69)	7794.61	24.546	28 (382)	29172.79	504.292

CONCLUSIONS

- Motivation → Simultaneity of the sport events
- Hypothesis → Approximation of the optimal solution diving the initial problem and solving each subproblem separately
- Contributions
  1. State of the art of clustering techniques with application to a given location-allocation problem
  2. State of the art on optimization methods
  3. Strategy to solve the immobile location-allocation problem
     - Dividing the problem using clustering
     - Applying optimization methods to every subproblem
  4. Clustering the search space
     - Clustering indices are useless to evaluate if a clustering is profitable to simplify an initial LA problem
     - Clustering the search space decrease the search time
     - Affinity propagation & k-means provide the best solutions.
  5. Optimization methods
     - Genetic algorithms needs a lot of memory resources
     - Simulated Annealing is the most efficient (best results in less amount of time)
     - The new neighborhood function improves the solution found by the algorithm
     - Initializing SA with the solution found by the individual method improves the performance

Clustering allows us to solve the problem

Clustering allows us to find a better solution

PAPERS

- F. Torrent, V. Muñoz, B. López. Exploring genetic algorithms and simulated annealing for immobile location-allocation problem. CCIA 2012.
- F. Torrent, V. Muñoz, B. López. An experimental analysis of clustering algorithms for supporting location-allocation. Submitted to AI 2012.

FUTURE WORK

- Develop an estimator of the customers' position just before the match
- Allow some permeability of the clusters' borders for the customers
- Use the true distance between bars and customers instead of the Euclidean distance
- Add other features to bars and customers (type of food, favorite team, etc.)
- Create a confidence index for each bar depending if they broadcast the assigned match
- Explore other partition techniques

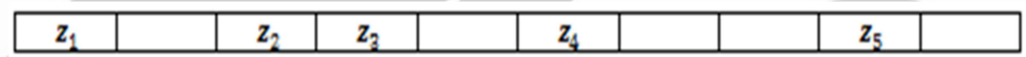
MOLTES GRÀCIES!!

Gràcies a:

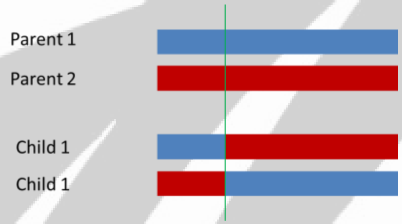
Beca UdG
Newronia S.L.
Grup eXiT

GA Based Clustering

- It determines the number of clusters
- Chromosome of length $L > N_{clusters}$



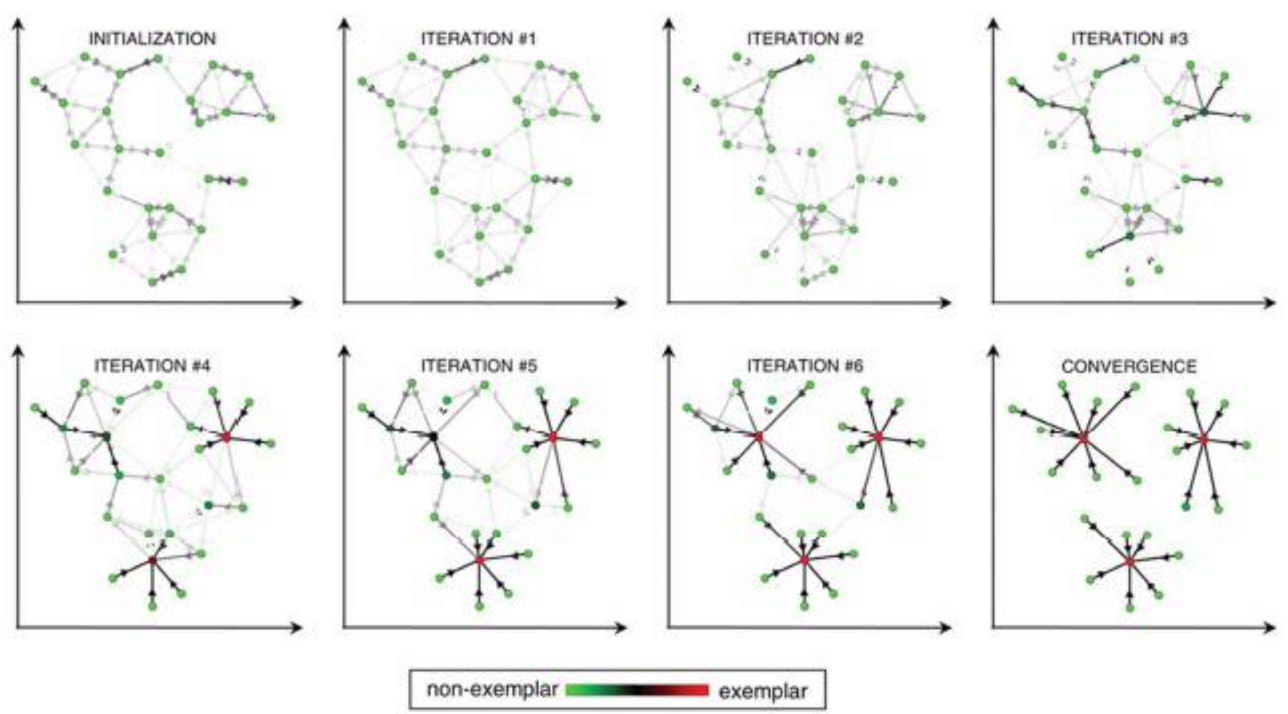
- Crossover
  - Single point crossover



- Mutation
  - $z_i = \begin{cases} z_i \cdot (1 \pm 2\delta) & z_i \neq 0 \\ \pm 2\delta & z_i = 0 \end{cases} \quad \delta \sim U(0,1)$
- Fitness
  - $Fitness = {}^{1}/_{DBI}$
- Selection
  - Roulette rule

AFFINITY PROPAGATION

- Elements exchange messages to vote the most representative ones
- It does not need any parameter



Clustering Results

Technique	Dataset 1 (459 bars)					Dataset 2 (1925)				
	CI	DBI	Num. clust.	Max clust.	Time (s)	CI	DBI	Num. clust.	Max clust.	Time (s)
Genetic	257.45	**0.664**	8	234	**34.574**	**1346.45**	**0.507**	18	**548**	**279.138**
Hierarchical	257.45	**0.664**	8	234	0.136	4745.74	0.451	48	395	69.871
K-means (empty clusters reassignment)	**1194.38**	0.462	128	55	17.336	18168.30	0.390	185	131	171.654
K-means (elements as centroids)	823.29	0.522	159	**53**	4.564	15825.53	0.342	834	**39**	101.654
Lloyd's alg.	628.10	0.473	170	44	18.081	44204.58	0.391	654	**39**	162.672
RG $D_{max} = 0.1$ km	419.74	**0.272**	172	73	**0.004**	**199033.78**	**0.100**	1082	**39**	**0.018**
RG $D_{max} = 0.2$ km	61.26	0.348	71	248	0.008	81860.09	0.174	770	102	0.045
RG $D_{max} = 0.5$ km	35.03	0.499	14	405	0.027	11297.91	0.216	401	257	0.098
RG $D_{max} = 1.0$km	**2.91**	0.466	2	**457**	0.043	7047.35	0.186	258	364	0.133
Affinity propagation	439.93	0.742	20	69	3.115	2819.31	0.565	28	382	49.415

GENETIC ALGORITHMS

Fitness of the final solution using different population sizes

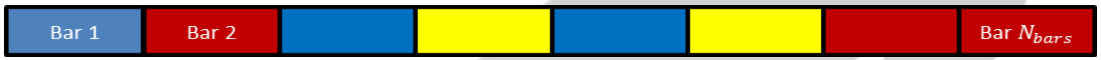
Population size	Number of facilities					
	8	18	48	50	72	127
5	**81.82**	199.53	**737.95**	**788.81**	**1128.17**	**1975.65**
10	82.66	**197.77**	750.44	810.24	1136.55	1977.84
25	**83.65**	199.98	760.38	822.99	1139.24	1986.52
50	83.61	200.50	757.23	818.76	1144.07	1985.82
100	**83.65**	**201.92**	759.06	**826.63**	1142.85	1984.38
150	**83.65**	201.76	**761.23**	821.06	**1145.83**	**1991.16**

Elapsed time using different population sizes

Population size	Number of facilities					
	8	18	48	50	72	127
5	**0.038**	**0.213**	**1.381**	**2.165**	**3.022**	**7.454**
10	0.056	0.304	3.524	3.439	6.625	21.242
25	0.169	0.730	9.165	9.632	18.948	52.218
50	0.394	1.822	21.424	23.238	40.701	101.783
100	0.696	4.163	42.833	46.181	85.762	223.688
150	**1.008**	**6.631**	**64.651**	**66.839**	**122.257**	**289.406**

Simulated Annealing

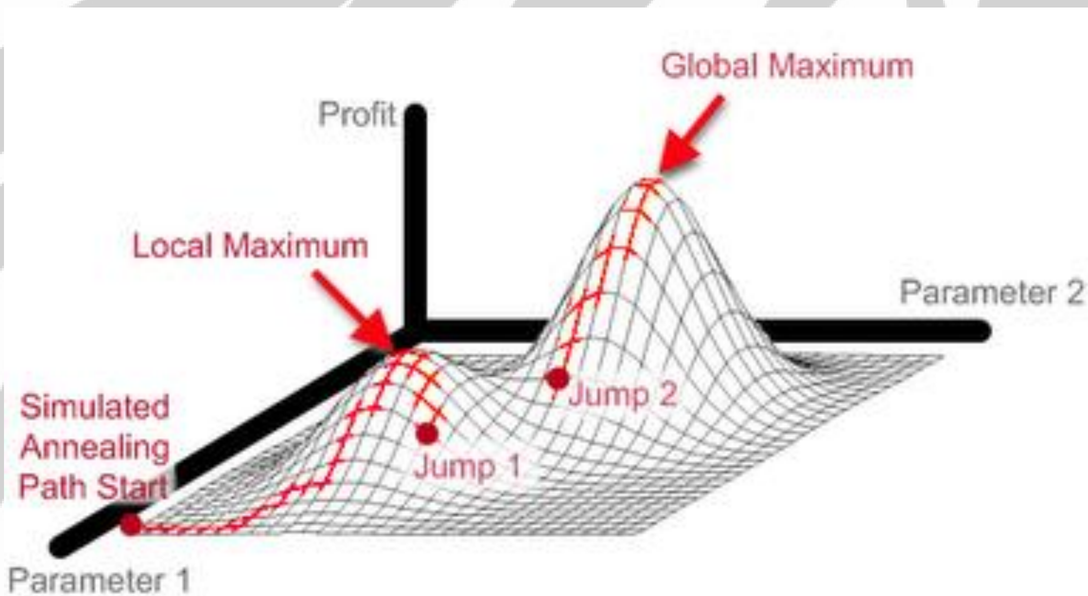
- Solution



- State selection

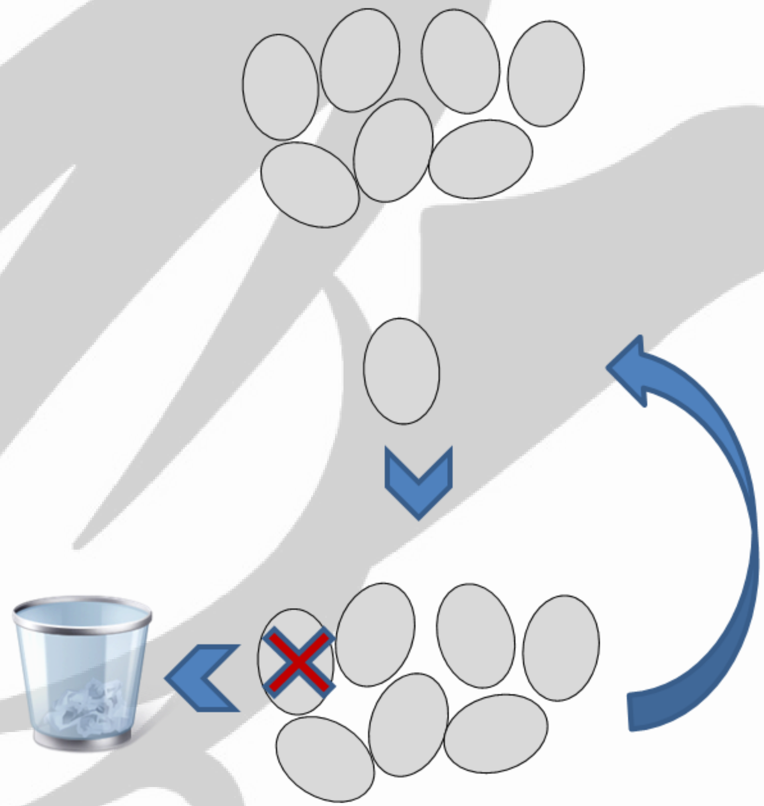
$$P\left(s' \mid E(s) < E(s')\right) = 1$$

$$P\left(s' \mid E(s) \geq E(s')\right) = e^{\frac{E(s') - E(s)}{T}}$$



CUCKOO SEARCH

1. Generate N random eggs (solutions)
2. Generate another random egg
3. Select one of the initial eggs
4. Substitute it by the new egg if it is better
5. Go back to step 2



CUCKOO SEARCH

- Non-coordinate search space → use of the previous neighborhood function
- The use of more nests does not imply a better solution but it implies more memory resources

Number of eggs	Number of facilities					
	8	18	48	50	72	127
2	72.02	206.45	719.79	646.06	1196.24	2013.67
5	72.02	**206.97**	722.99	645.30	1195.67	2008.25
10	72.02	201.75	718.47	**646.54**	1196.38	**2019.30**
15	**72.08**	202.76	720.30	645.75	1185.14	2003.64
20	69.20	204.60	719.27	644.79	1188.04	2017.85
25	71.77	195.47	**724.88**	646.29	**1197.16**	2008.37

Clustering Indices

- Calinski index

$$C(k)=\frac{B(k)/k-1}{W(g)/n-k}$$

$$B(k)=\sum_{i=1}^{k} n_i \left\|\mathbf{z}_i-\mathbf{z}\right\|^2 \qquad W(k)=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\|\mathbf{x}_j-\mathbf{z}_i\right\|^2 \qquad z=\frac{\sum_{i=1}^{n}\mathbf{x}_i}{n}$$

- Davies-Bouldin index

$$DB(k)=\frac{1}{k}\sum_{i=1}^{k} R_{i,qt}$$

$$R_{i,qt}=\max_{j,j\neq i}\left\{\frac{S_{i,q}+S_{j,q}}{d_{ij,t}}\right\} \qquad d_{ij,t}=\sqrt[t]{\sum_{s=1}^{P}\left|z_{is}-z_{js}\right|^t}=\mathbf{z}_i-\mathbf{z}_{jt} \qquad S_{i,q}=\sqrt[q]{\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_j-\mathbf{z}_{i2}^{\,q}}$$