Approximately Pareto-optimal Solutions for Bi-Objective k-Clustering Problems

Anna Arutyunova, **Jan Eube**, Heiko Röglin,
Melanie Schmidt, Sarah Sturm, Julian Wargalla

Clustering

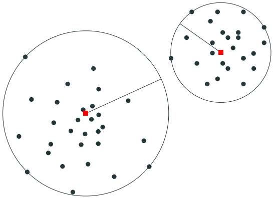
Given: Metric space $(\mathcal{P}, d)$ with $|\mathcal{P}| = n$.

k-**clustering:** Choose k centers $C \subset \mathcal{P}$ and assignment $\sigma\colon \mathcal{P} \to C$ optimizing

k-center: $\mathsf{rad}(C, \sigma) = \max_{p \in P} d(p, \sigma(p))$ **k-means:** $\mathsf{mean}(C, \sigma) = \sum_{p \in P} d(p, \sigma(p))$ **k-separation:** $\mathsf{sep}(C, \sigma) = \min_{\substack{p,q \in P \\ \sigma(p) \neq \sigma(q)}} d(p, q)$

Given: Metric space $(\mathcal{P}, d)$ with $|\mathcal{P}| = n$.

k-**clustering:** Choose k centers $C \subset \mathcal{P}$ and assignment $\sigma\colon \mathcal{P} \to C$ optimizing



k-center: $\mathsf{rad}(C, \sigma) = \max_{p \in P} d(p, \sigma(p))$

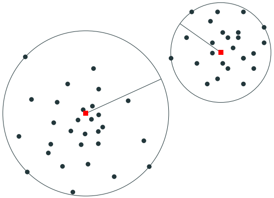
k-means: $\mathsf{mean}(C, \sigma) = \sum_{p \in P} d(p, \sigma(p))$

k-separation: $\mathsf{sep}(C, \sigma) = \min_{\substack{p,q \in P \\ \sigma(p) \neq \sigma(q)}} d(p, q)$

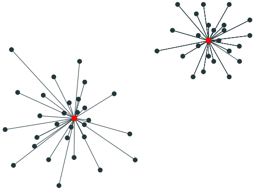
Clustering

Given: Metric space $(\mathcal{P}, d)$ with $|\mathcal{P}| = n$.

k-**clustering:** Choose k centers $C \subset \mathcal{P}$ and assignment $\sigma \colon \mathcal{P} \to C$ optimizing



k-center: $\mathsf{rad}(C, \sigma) = \max_{p \in P} d(p, \sigma(p))$



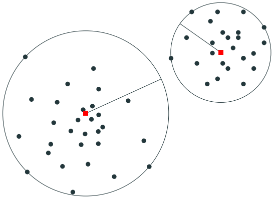
k-means: $\mathsf{mean}(C, \sigma) = \sum_{p \in P} d^2(p, \sigma(p))$

k-separation: $\mathsf{sep}(C, \sigma) = \min_{\substack{p,q \in P \\ \sigma(p) \neq \sigma(q)}} d(p, q)$

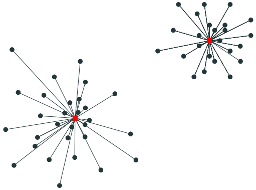
Clustering

Given: Metric space $(\mathcal{P}, d)$ with $|\mathcal{P}| = n$.

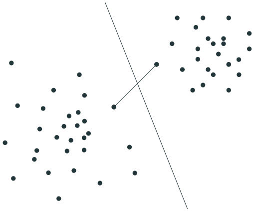
k-**clustering:** Choose k centers $C \subset \mathcal{P}$ and assignment $\sigma\colon \mathcal{P} \to C$ optimizing



k-center: $\mathsf{rad}(C, \sigma) = \max_{p \in P} d(p, \sigma(p))$

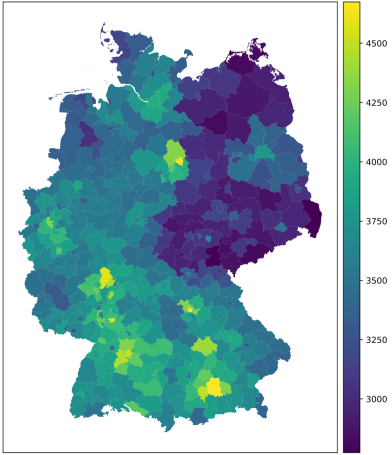


k-means: $\mathsf{mean}(C, \sigma) = \sum_{p \in P} d^2(p, \sigma(p))$



k-separation: $\mathsf{sep}(C, \sigma) = \min_{\substack{p,q \in P \\ \sigma(p) \neq \sigma(q)}} d(p, q)$

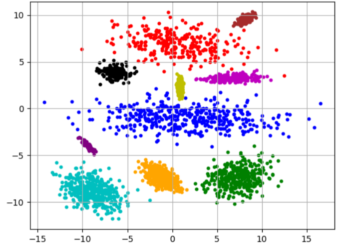
What if we want to consider multiple metrics?



Average income of the German districts in Euro[1]

[1]Source: Bundesagentur für Arbeit

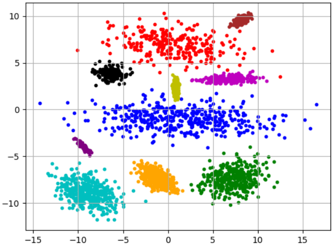
What if we want to consider multiple objective functions?



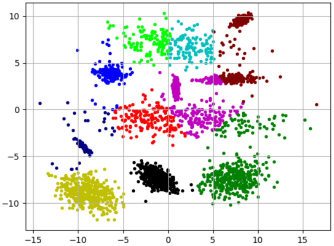
The ground truth.[2]

[2]Data Set generated by Julia Handl and Joshua D. Knowles, 2007

What if we want to consider multiple objective functions?



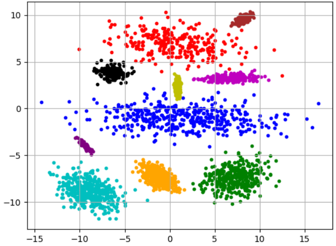
The ground truth.[2]



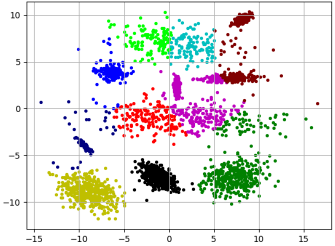
Clustering by k-means++

[2] Data Set generated by Julia Handl and Joshua D. Knowles, 2007

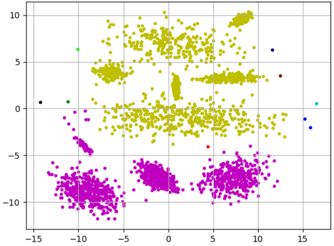
What if we want to consider multiple objective functions?



The ground truth.[2]



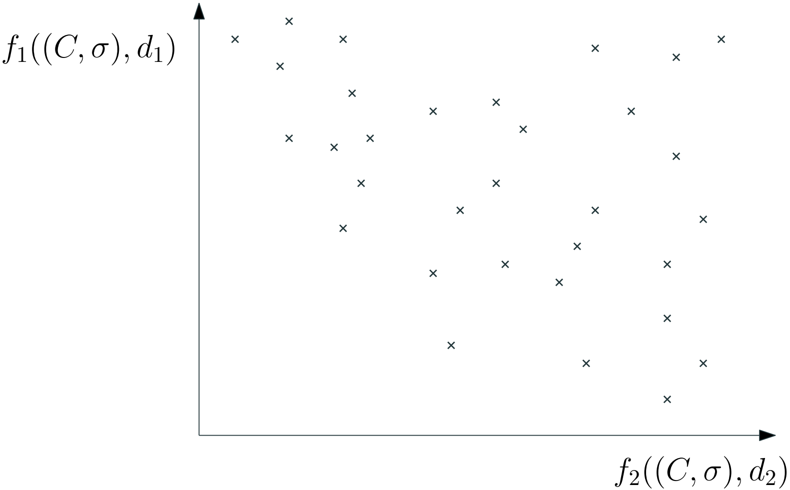
Clustering by *k*-means++



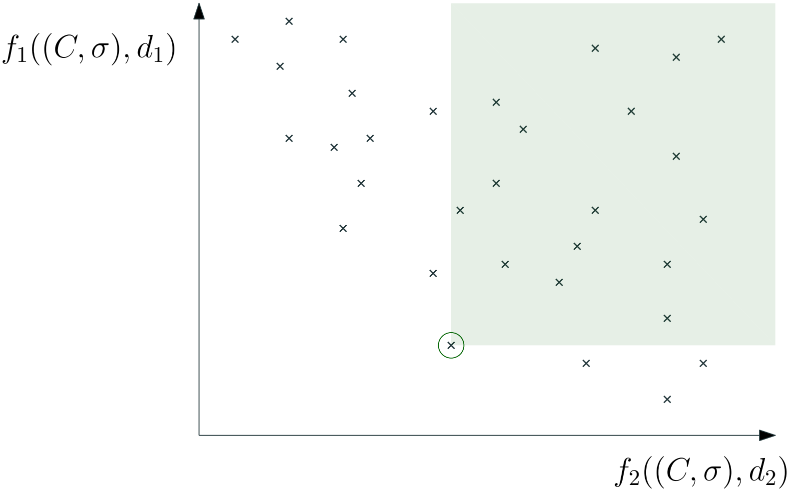
Optimum *k*-separation clustering

[2] Data Set generated by Julia Handl and Joshua D. Knowles, 2007

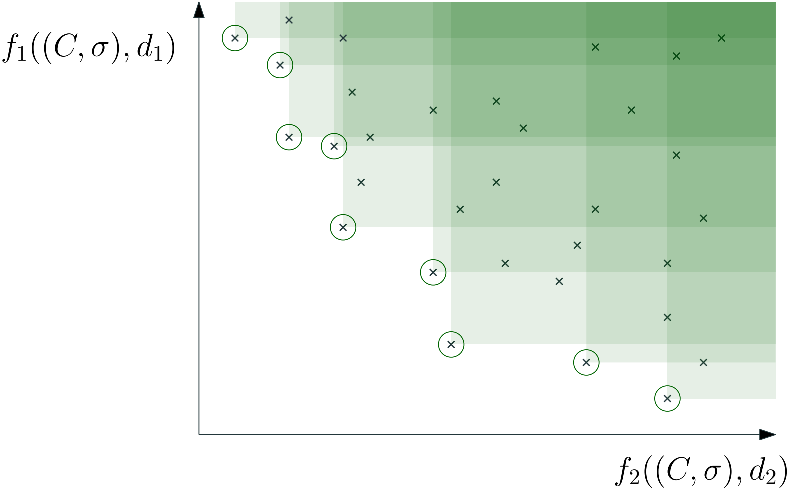
Pareto Sets



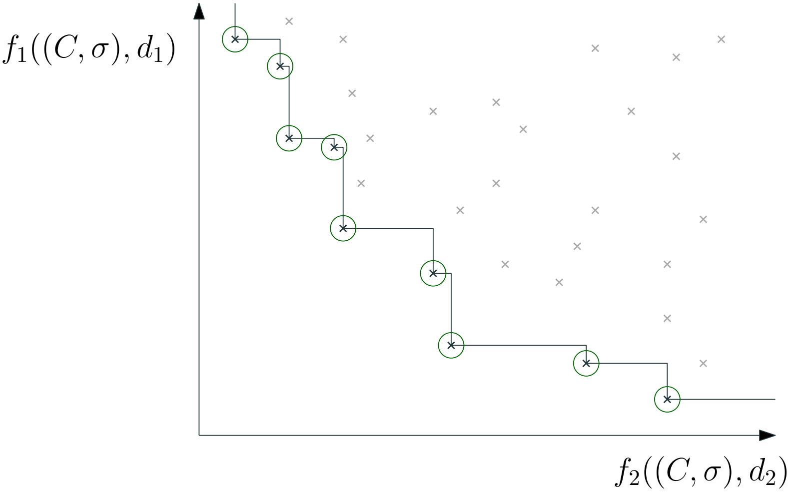
Pareto Sets



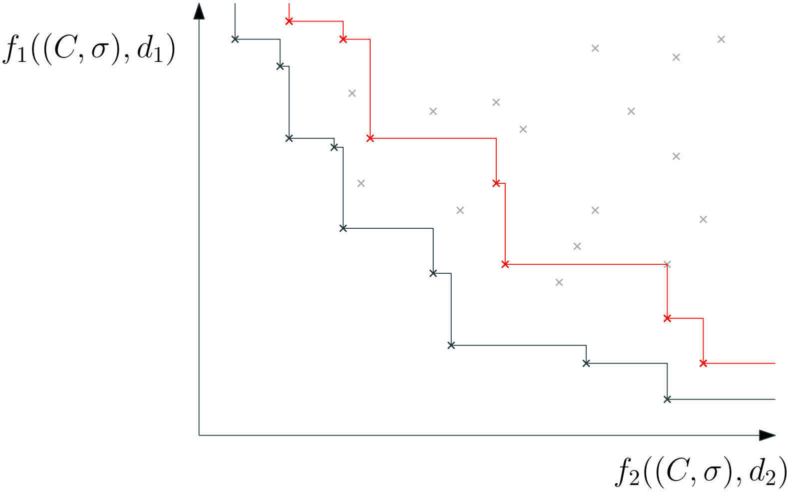
Pareto Sets



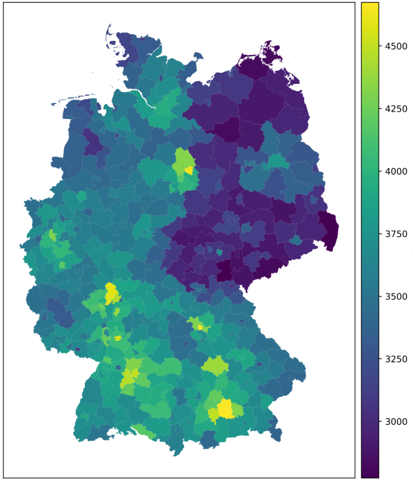
Pareto Sets



Approximate Pareto Set

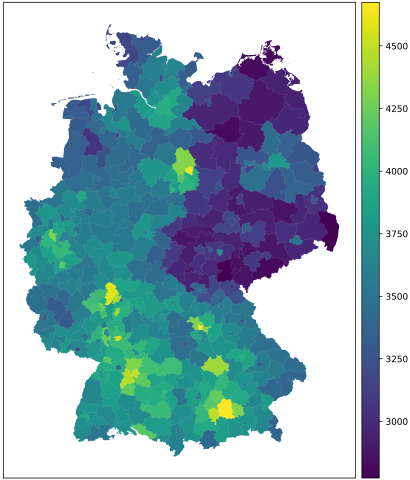


Experimental Evaluation

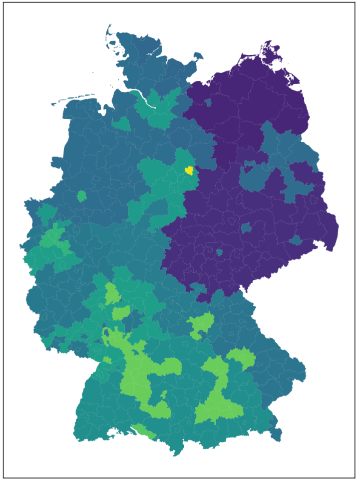


Average income of the German districts in Euro

Experimental Evaluation



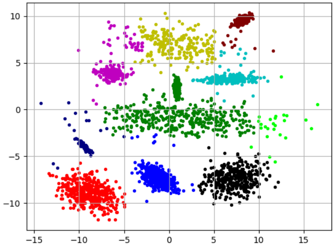
Average income of the German districts in Euro



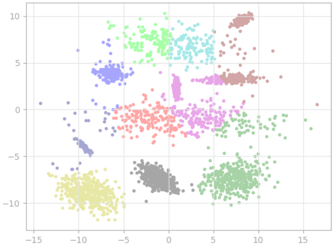
A good Pareto solution

Experimental Evaluation

Optimum *k*-separation clustering



The best Pareto Solution



Clustering by *k*-means++

Theoretical Results

	k-center	k-diameter	k-separation	k-median	k-means
k-center	(2,2)	(2,2)	(2, 1)	$(9,\ 6+\epsilon)$[1]	$(9{,}54+\epsilon)$
k-diameter	-	(2,2)	(2, 1)	$(18{,}6+\epsilon)$	$(18{,}54+\epsilon)$
k-separation	-	-	(1, 1)	$(1{,}2+\delta_1)$[2]	$(1,\ 4+4\sqrt{\delta_2}+\delta_2)$[2]
k-median	-	-	-	$2.675+\epsilon$[3]	$9+\epsilon$[3]
k-means	-	-	-	-	$9+\epsilon$[3]

[1] Alamdari & Shmoys, 2017 also provided a $(4, 8)$ approximation algorithm

[2] δ_1 and δ_2 refer to the currently best single-objective approximation factors respectively.

[3] Calculates only an approximate convex Pareto Set, the approximation is the same for both objectives

Theoretical Results

	k-center	k-diameter	k-separation	k-median	k-means
k-center	(2,2)	(2,2)	(2, 1)	$(9, 6+\epsilon)$[1]	$(9, 54+\epsilon)$
k-diameter	-	(2,2)	(2, 1)	$(18, 6+\epsilon)$	$(18, 54+\epsilon)$
k-separation	-	-	(1, 1)	$(1, 2+\delta_1)$[2]	$(1, 4+4\sqrt{\delta_2}+\delta_2)$[2]
k-median	-	-	-	$2.675+\epsilon$[3]	$9+\epsilon$[3]
k-means	-	-	-	-	$9+\epsilon$[3]

[1] Alamdari & Shmoys, 2017 also provided a $(4, 8)$ approximation algorithm

[2] δ_1 and δ_2 refer to the currently best single-objective approximation factors respectively.

[3] Calculates only an approximate convex Pareto Set, the approximation is the same for both objectives

Thank you for your attention!