Space-filling location selection

Michela Bia & Philippe Van Kerm

CEPS/INSTEAD, Luxembourg

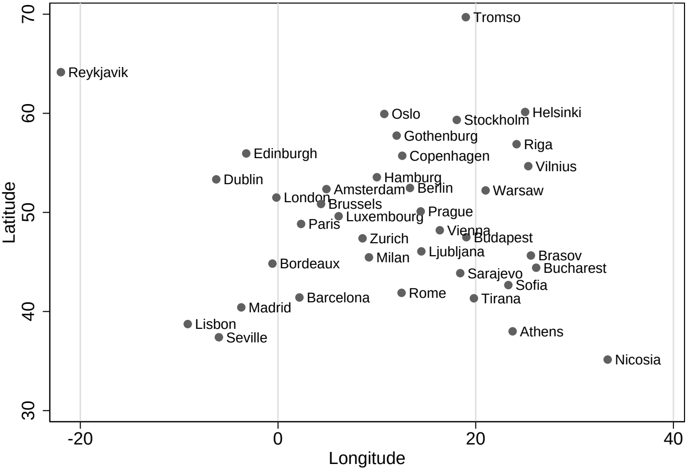
philippe.vankerm@ceps.lu

2014 London Stata Users Group meeting

September 11–12 2014, Cass Business School, London

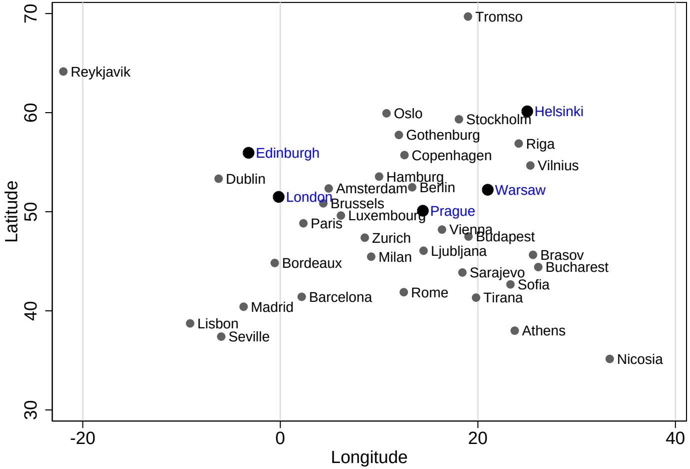
Location selection

Where to place 5, say, sensors within a set of N candidate locations ?



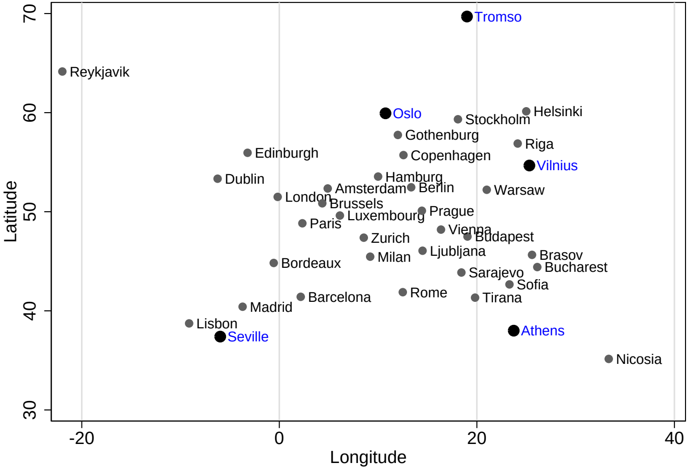
Location selection

Here?



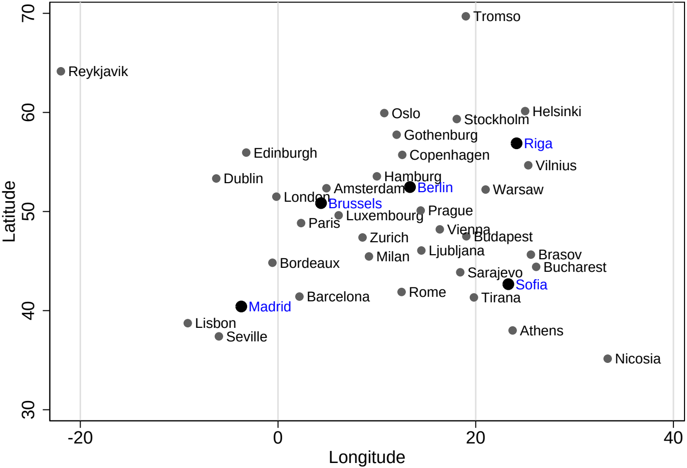
Location selection

Or here?



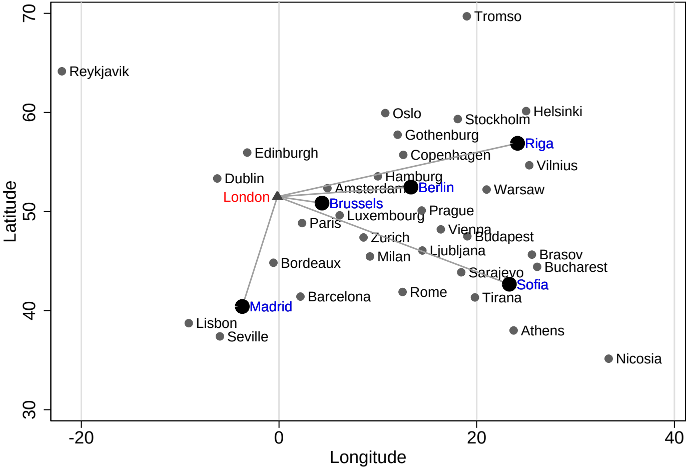
Location selection

Or here? (One of $\mathcal{C}$ possibilites)



Minimize distances to the set

How far is London from the set of 5 cities (the ‘design’ $\mathcal{D}_n$, $n = 5$)?



Geometric coverage criterion (1)

Coverage of a location by a set of locations

The coverage of single location

Distance between a given location $\mathbf{x}$ and a design $\mathcal{D}_n$ is

$$d_p(\mathbf{x}, \mathcal{D}_n) = \left(\sum_{\mathbf{y} \in D_n} ||\mathbf{x} - \mathbf{y}||^p \right)^{\frac{1}{p}} \qquad (1)$$

with $p < 0$.

$d_p(\mathbf{x}, \mathcal{D}_n)$ measures how well the design $\mathcal{D}_n$ ‘covers’ the location $\mathbf{x}$.

When $p \longrightarrow -\infty$, $d_p(\mathbf{x}, \mathcal{D}_n)$ tends to the shortest Euclidian distance between $\mathbf{x}$ and a point in $\mathcal{D}_n$.

See Johnson et al. (1990), Royle & Nychka (1998).

Geometric coverage criterion (2)

Coverage of multiple locations

The coverage of the multiple locations

The coverage by $\mathcal{D}_n^*$ of *multiple* locations given by (q-power) mean of the individual coverages:

$$C_{p,q}(\mathcal{C}, \mathcal{D}_n) = \left(\sum_{\mathbf{x} \in \mathcal{C}} d_p(\mathbf{x}, \mathcal{D}_n)^q \right)^{\frac{1}{q}} \qquad (2)$$

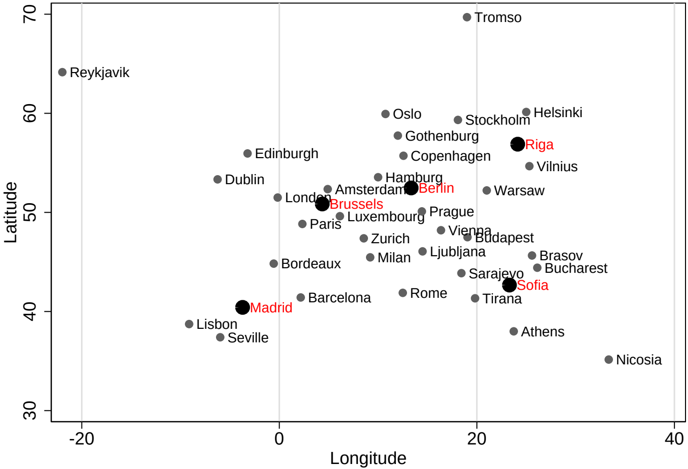
High q gives high penalty to the large distances.

Optimal design $\mathcal{D}_n^*$

An ‘optimal design’ of size n given parameters p and q is the combination of n locations that minimizes $C_{p,q}(\mathcal{C}, \mathcal{D}_n)$.

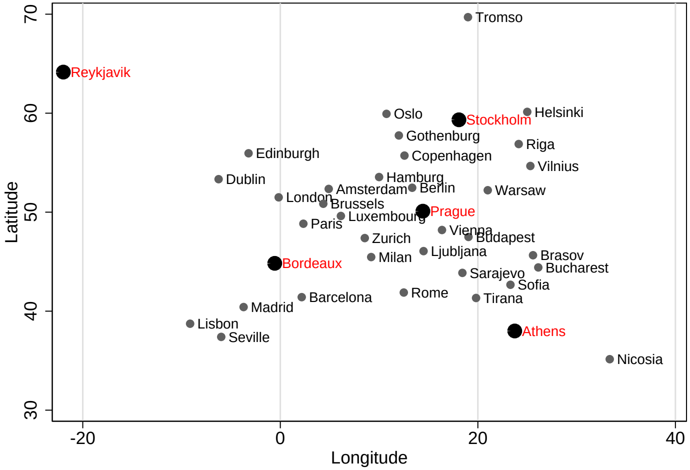
Optimal subset with $p = -5$ and $q = 1$

Arithmetic mean: short and long distances weighted equally



Optimal subset with $p = -5$ and $q = 5$

Increasing q implies higher penalty for long distances



Numerical optimization: point-swapping algorithm

Brute-force calculation of optimal design computationally prohibitive!

Royle & Nychka (1998) point-swapping algorithm:

1. Start from a random initial design $\mathcal{D}_n^0$
2. take a location from initial design and swap with the candidate point leading to greatest improvement in coverage (if any)
3. iterate for all points in design until no swap improves coverage

Speed improvement: in step 2, search only among k nearest candidate locations

Algorithm always converges to a solution, but not necessarily globally optimal with nearest neighbour swaps.

Repeat optimization with alternative initial random subsets.

Imposing constraints

It is straightforward to impose constraints:

- force inclusion of particular locations into the design
- exclude particular locations from the design (but include them in calculations of coverage)
- separate the candidate set from the set to be covered

Higher dimensional data and grid selection

- algorithm is not limited to two dimensional and/or geographic data
- requires common metric for geometric distance to be meaningful: normalization, scaling
- key application: grid selection for non-parametric regression models in large samples (e.g., multivariate kernel density estimation, local regression)—see below.

The **spacefill** command: Syntax

Syntax

spacefill *varlist* [*weight*] [*if*] [*in*] , [ndesign(#)
design0(*varlist*) fixed(*varname*) exclude(*varname*) p(#) q(#)
nnfrac(#) nnpoints(#) nruns(#) standardize standardize2
standardize3 sphericize ranks generate(*newvar*)
genmarker(*newvar*) noverbose]

fweight, aweight, pweight and iweight are allowed; see [U] **11.1.6 weight – Weights**.

varlist and the [*if*] or [*in*] clauses identify the data from which the optimal subset is selected.

The `spacefill` command: Key options

`ndesign(#)` specifies *n*, the size of the design. Default is 4.

`design0(`*varlist*`)` identifies a (set of) initial designs

`fixed(`*varname*`)` identifies observations that are included in all designs

`exclude(`*varname*`)` identifies observations excluded from all designs

`p(#)` and `q(#)` specify the *p* and *q* parameters

`nnfrac(#)` specifies the fraction of data to consider as nearest neighbours in the point-swapping iterations. Default is 0.50.

`nruns(#)` sets the number of independent runs performed on alternative random initial designs.

`standardize`, `standardize2`, `standardize3`, `sphericize`, `ranks` standardize all variables in *varlist* before calculating distances between observations.

Example 1: City selection

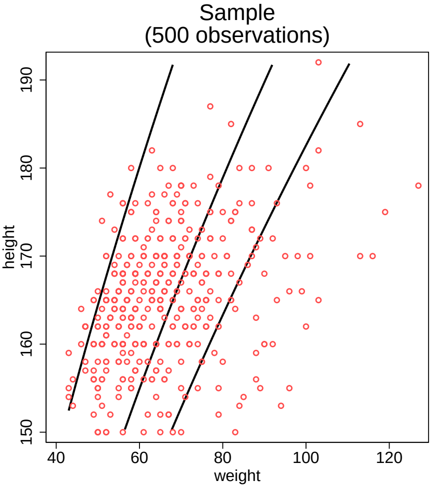
```
use eu-cities.dta, clear
spacefill lat lon , nd(5) nnfrac(1) genmark(select1)
spacefill lat lon , nd(5) q(5) nnfrac(1)
genmark(select2)
spacefill lat lon , nd(5) nnfrac(.5) nruns(10)
```

Example 2: Grid selection for local regression

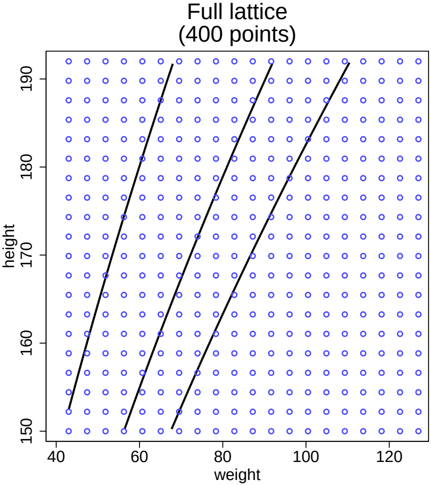
- Data on height, weight and wage of a sample of working women in Luxembourg
- Any relationship between stature and wages?
- Non-parametric relationship estimated via local regression (more flexible than regressing wages on body mass index)
- No need to estimate the expected wage at *all* height and weight combination observed in the data: estimates on a grid is enough
- but a simple regular, rectangular grid not useful since some combinations of height and weight are never observed

$\Longrightarrow$ use space-filling to select points on the grid that best cover the sample

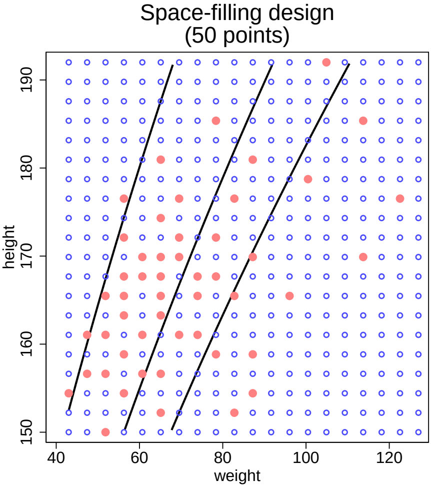
Example 2: Grid selection in local regression



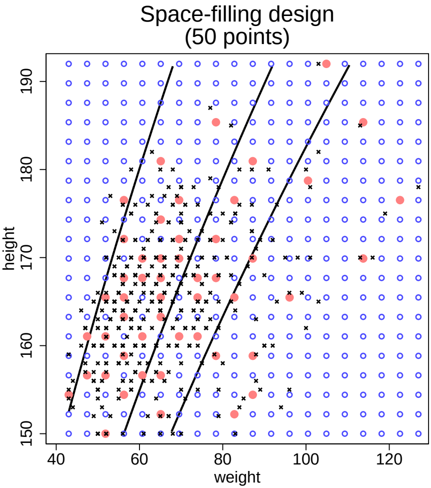
Example 2: Grid selection in local regression



Example 2: Grid selection in local regression



Example 2: Grid selection in local regression

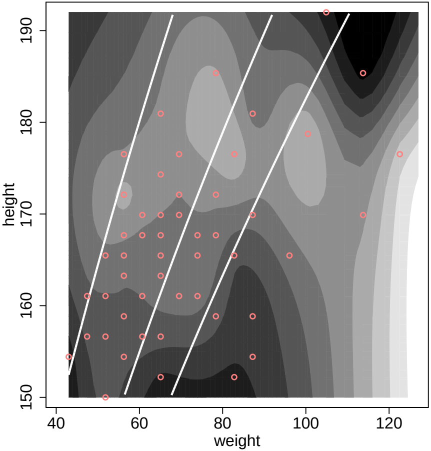


Example 2: Grid selection

```
set obs 20
range height 150 192
range weight 43 127
fillin height weight
gen byte sample2 = 0
save gridheiwei.dta , replace
use "Height_Weight_Wage.dta", clear
gen byte sample2 = 1
append using gridheiwei
spacefill height weight [iw=sample2], exclude(sample2)
nd(50) nr(10) nnfrac(.20) standardize2
```

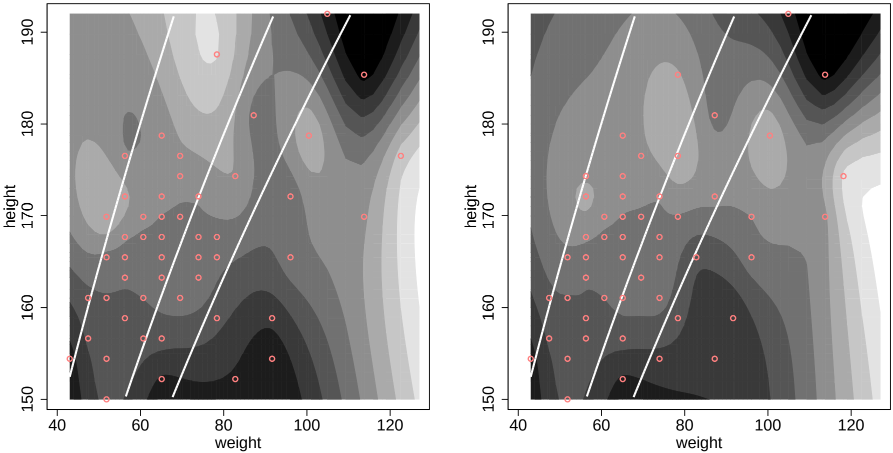
Example 2: Grid selection in local regression

local regression estimates on grid + contour plot display



Example 2: Grid selection in local regression

Variations in the selected design



References

- Bia M. and Van Kerm P. Space-filling location selection. Stata Journal, forthcoming. Working paper available at `http://ideas.repec.org/p/irs/cepswp/2013-17.html`.
- Furrer, R., D. Nychka, and S. Sain. 2013. fields: Tools for spatial data. R package version 6.7.6, 2013-04-21. http://CRAN.R-project.org/package=fields
- Johnson, M. E., L. M. Moore, and D. Ylvisaker. 1990. Minimax and maximin distance designs. Journal of Statistical Planning and Inference 26(2): 131–148.
- Royle, J. A. and D. Nychka. 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Computers & Geosciences 24(5): 479–488.

Acknowledgements

This work is part of the project *“Estimation of direct and indirect causal effects using semi-parametric and non-parametric methods”* supported by the Luxembourg “Fonds National de la Recherche” cofunded under the Marie Curie Actions of the European Commission (FP7-COFUND) and of the project *“Information and Wage Inequality”* supported by the Luxembourg Fonds National de la Recherche (contract C10/LM/785657).