Urban Data Analytics

–

Seeing cities through the lens of mobile phone data

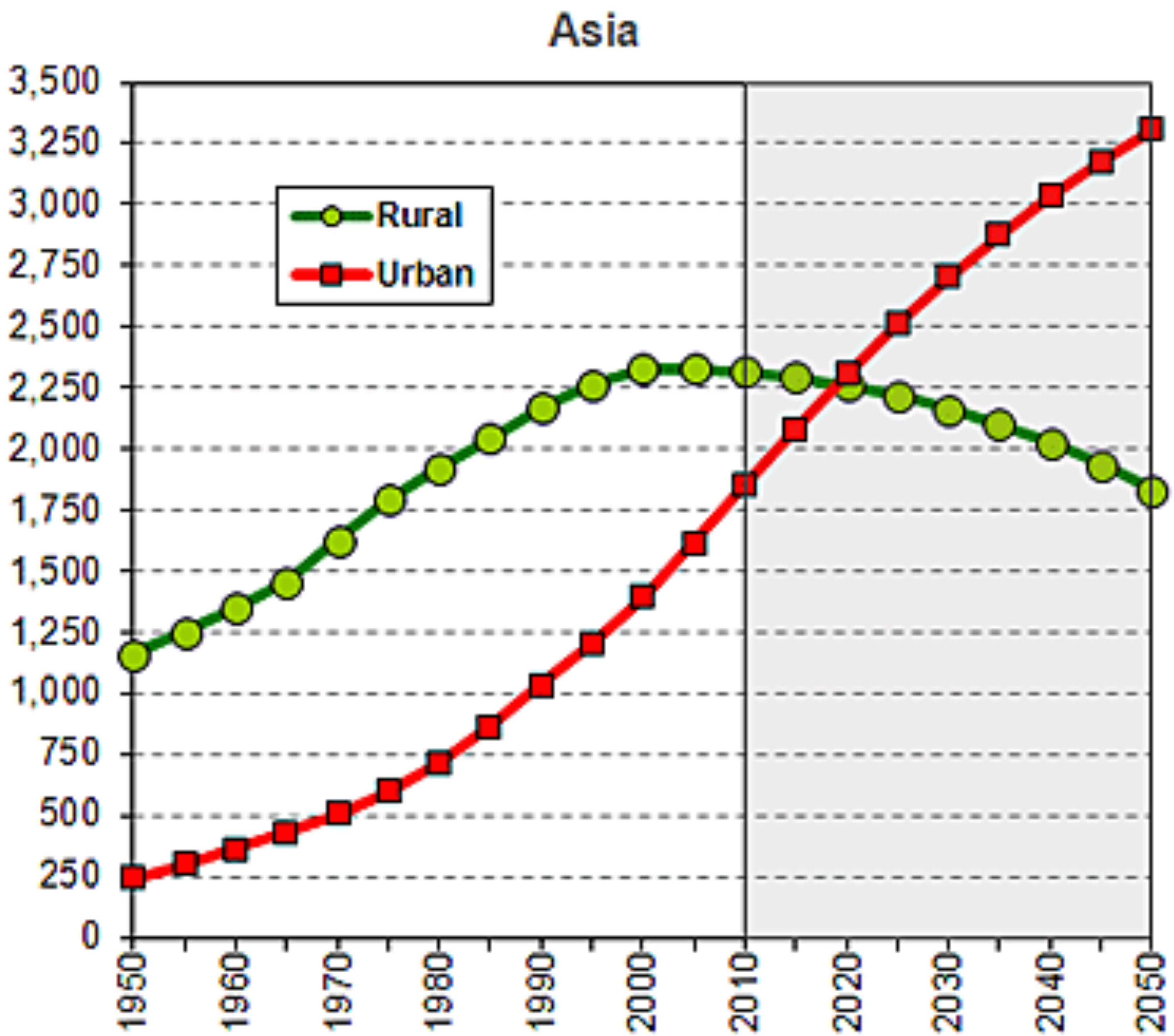
Markus Schläpfer
ETH Future Cities Laboratory

NTU Winter School on Complexity
March 11, 2016


NOVARTIS

Ever-increasing complexity of cities

A. Population growth and mass urbanization



UN Population Division, 2012

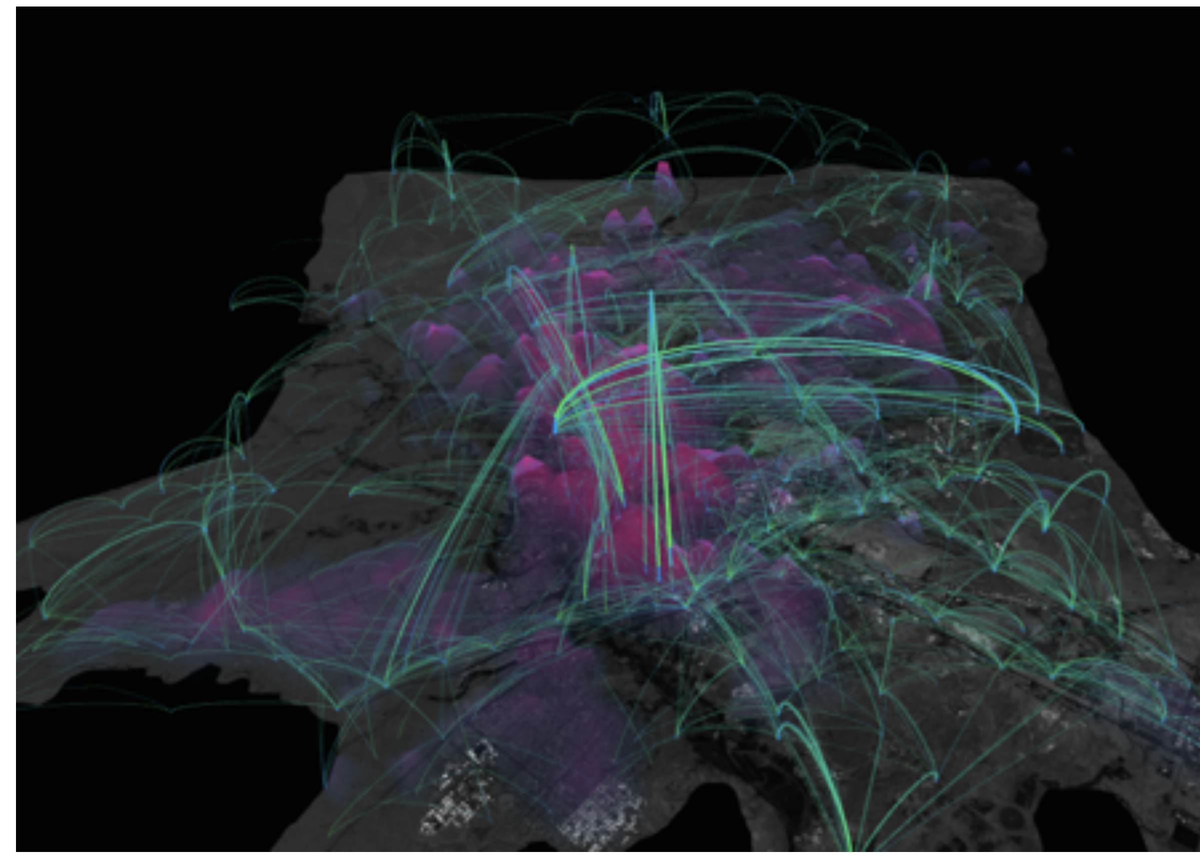
Ever-increasing complexity of cities

B. New forms of urban organization

„Monocentricity"

Univ. Munster

„Polycentricity“



MIT/Senseable City Lab, Kael Greco

Increasing uncertainties in urban planning and design

- Urban mobility
- Infrastructure design
- Social sustainability
  (social segregation, job accessibility)
- …

Urgent need for a quantitative *understanding* of cities

Growing availability of human activity data

Growing availability of human activity data

- Mobile phone data
- Smart card data from public transportation
- GPS traces from vehicular devices
- Location-based social networks

  (Foursquare, Twitter, Flickr, Running Apps, etc.)
- User-generated mapping projects (OpenStreetMap)

Growing availability of human activity data

- Mobile phone data

Mobile phone data - exemplary data sources

- Open data
    - *Italy* - Telecom Italia Open BigData Initiative
      http://theodi.fbk.eu/openbigdata

- Big data research competitions
    - *Ivory Coast* - Orange D4D Challenge 2013
      http://www.d4d.orange.com/en/Accueil
    - *Senegal* - Orange D4D Challenge 2015
      http://www.d4d.orange.com/en/Accueil
    - *Italy* - Telecom Italia BigData Challenge 2015
      http://www.telecomitalia.com/tit/en/bigdatachallenge.html

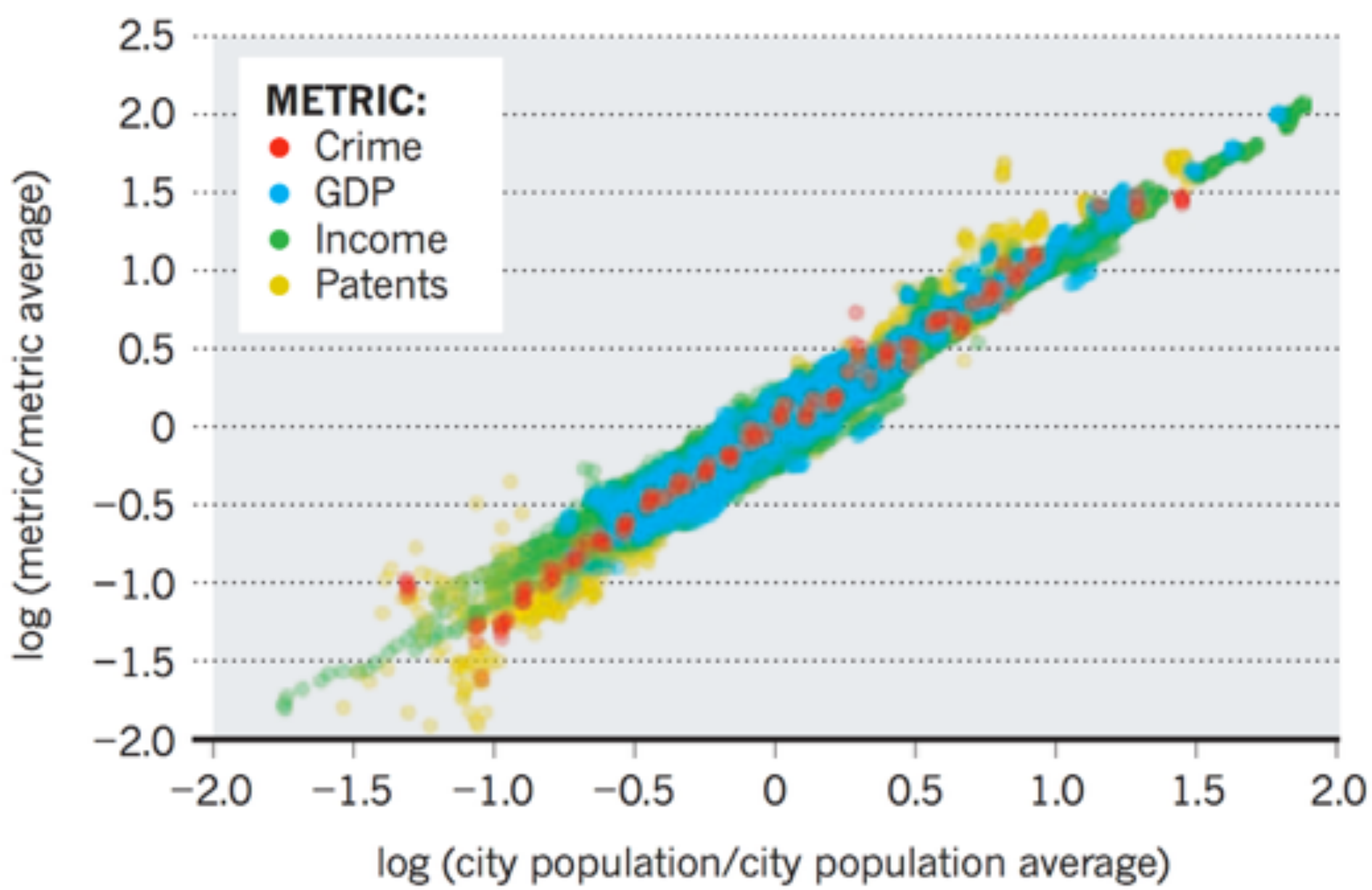
- (Telco providers and data analytics companies)

Two exemplary applications

1. Measuring human interactions in cities
2. Quantifying the movement of people in cities

Part I: Human interactions

Super-linear scaling of socio-economic quantities



L.M.A. Bettencourt & G.B. West, Nature (2010)

$$Y \propto N^{\beta} \qquad \beta \approx 1.15 > 1$$

Y Socio-economic quantity

N City population size

β Scaling exponent

Network of human interactions as a unifying mechanism?

Growth, innovation, scaling, and the pace of life in cities

Luís M. A. Bettencourt*†, José Lobo‡, Dirk Helbing§, Christian Kühnert§, and Geoffrey B. West*¶

*Theoretical Division, MS B284, Los Alamos National Laboratory, Los Alamos, NM 87545; ‡Global Institute of Sustainability, Arizona State University, P.O. Box 873211, Tempe, AZ 85287-3211; §Institute for Transport and Economics, Dresden University of Technology, Andreas-Schubert-Strasse 23, D-01062 Dresden, Germany; and ¶Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

Edited by Elinor Ostrom, Indiana University, Bloomington, IN, and approved March 6, 2007 (received for review November 19, 2006)

Humanity has just crossed a major landmark in its history with the majority of people now living in cities. Cities have long been known to be society's predominant engine of innovation and

The increasing concentration of people in cities presents both opportunities and challenges (9) toward future scenarios of sustainable development. On the one hand, cities make possible

PHYSICAL REVIEW E **79**, 016115 (2009)

Superlinear scaling for innovation in cities

Samuel Arbesman*
Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, Massachusetts 02115, USA

Jon M. Kleinberg†
Computer Science, Cornell University, Ithaca, New York 14853, USA

Steven H. Strogatz‡
Theoretical and Applied Mechanics, Cornell University, Ithaca, New York 14853, USA

(Received 29 September 2008; published 30 January 2009)

Superlinear scaling in cities, which appears in sociological quantities such as economic productivity and creative output relative to urban population size, has been observed, but not been given a satisfactory theoretical explanation. Here we provide a network model for the superlinear relationship between population size and innovation found in cities, with a reasonable range for the exponent.

nature COMMUNICATIONS

ARTICLE

Received 4 Dec 2012 | Accepted 30 Apr 2013 | Published 4 Jun 2013

DOI: 10.1038/ncomms2961

Urban characteristics attributable to density-driven tie formation

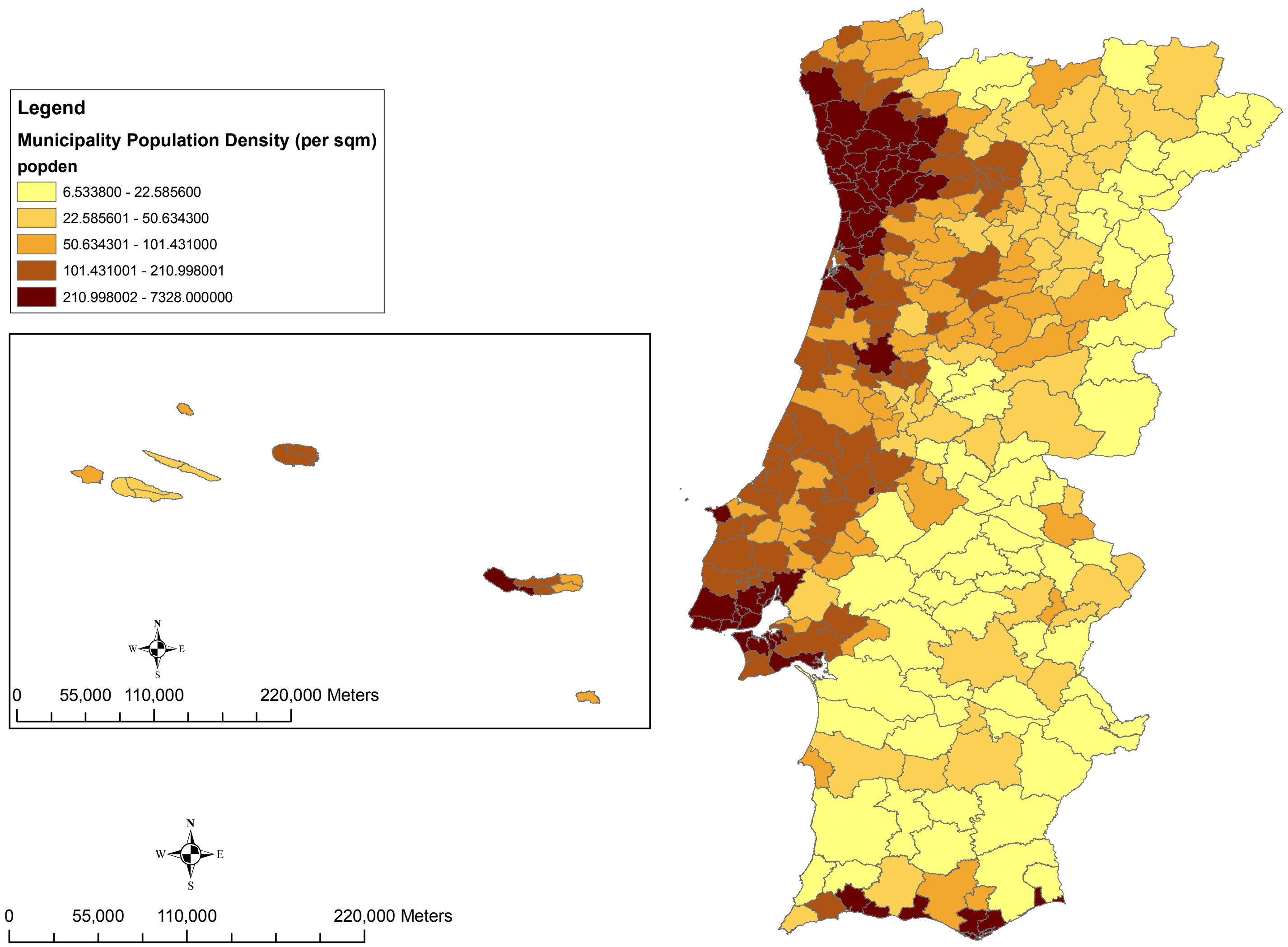
Wei Pan[1], Gourab Ghoshal[1,†], Coco Krumme[1], Manuel Cebrian[1,2,3] & Alex Pentland[1]

The Origins of Scaling in Cities

Luís M. A. Bettencourt

Despite the increasing importance of cities in human societies, our ability to understand them scientifically and manage them in practice has remained limited. The greatest difficulties to any scientific approach to cities have resulted from their many interdependent facets, as social, economic, infrastructural, and spatial complex systems that exist in similar but changing forms over a huge range of scales. Here, I show how all cities may evolve according to a small set of basic principles that operate locally. A theoretical framework was developed to predict the

Lets look into the data!



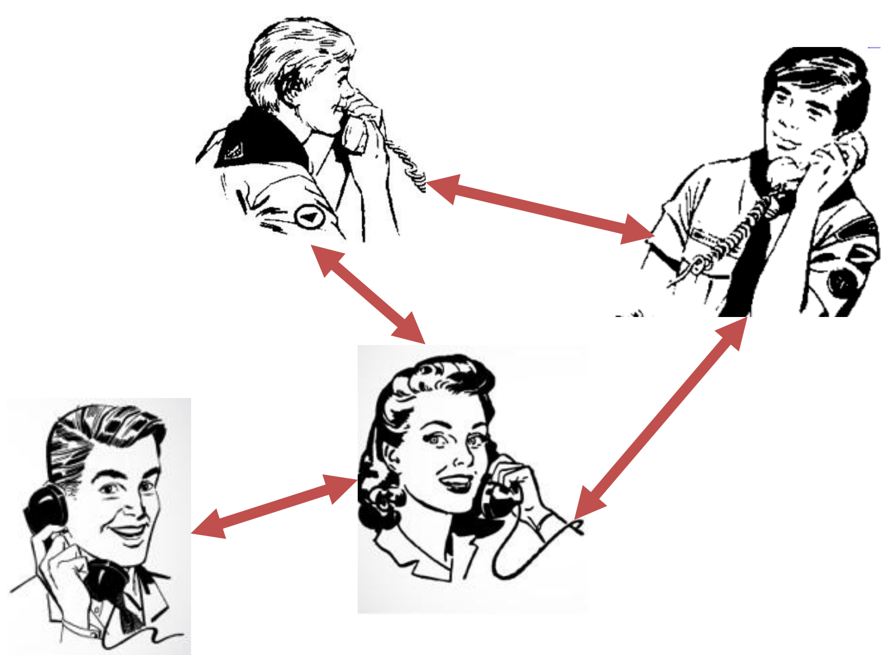
Several millions of anonymized call detail records (CDRs) from Portugal for a period of ≈15 months

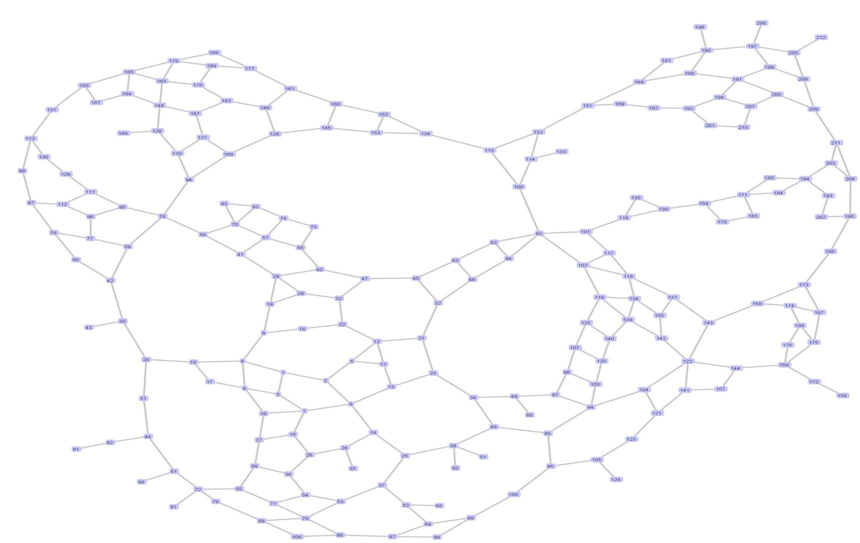
Call detail records (CDRs)

- Anonymized ID (surrogate number) of the caller
- Anonymized ID of the callee
- Start time of the call
- Duration of the call
- The locations of the antennas routing the call

Inferring the interaction network

Mobile phone user	→	Node
Reciprocal call between two users	→	Link





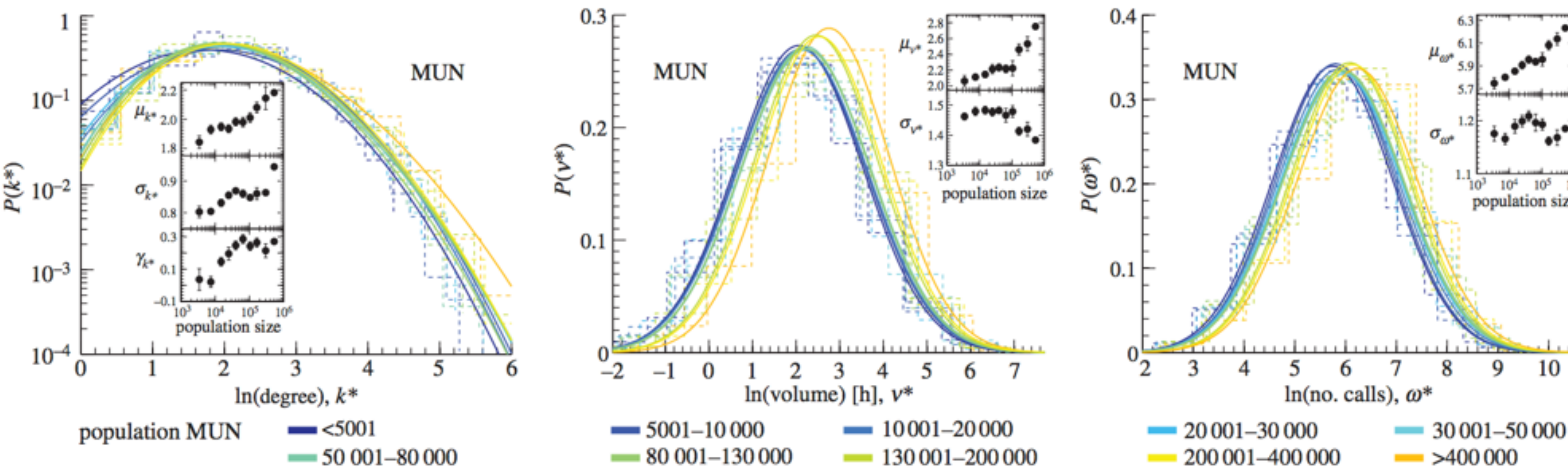
Portugal data:
1.6 Mio nodes
6.8 Mio links

Individual-based interaction distributions

degree

call volume

number of calls



Schläpfer M. et al., J Royal Soc Interface, 11(98): 20130789, 2014

Scaling of human interactions

$$Y \propto N^{\beta}$$

Expectation

Upper bound: $\beta < 2$

More realistic: $\beta = 1 + \dfrac{\log A}{\log(N_1/N_2)}$

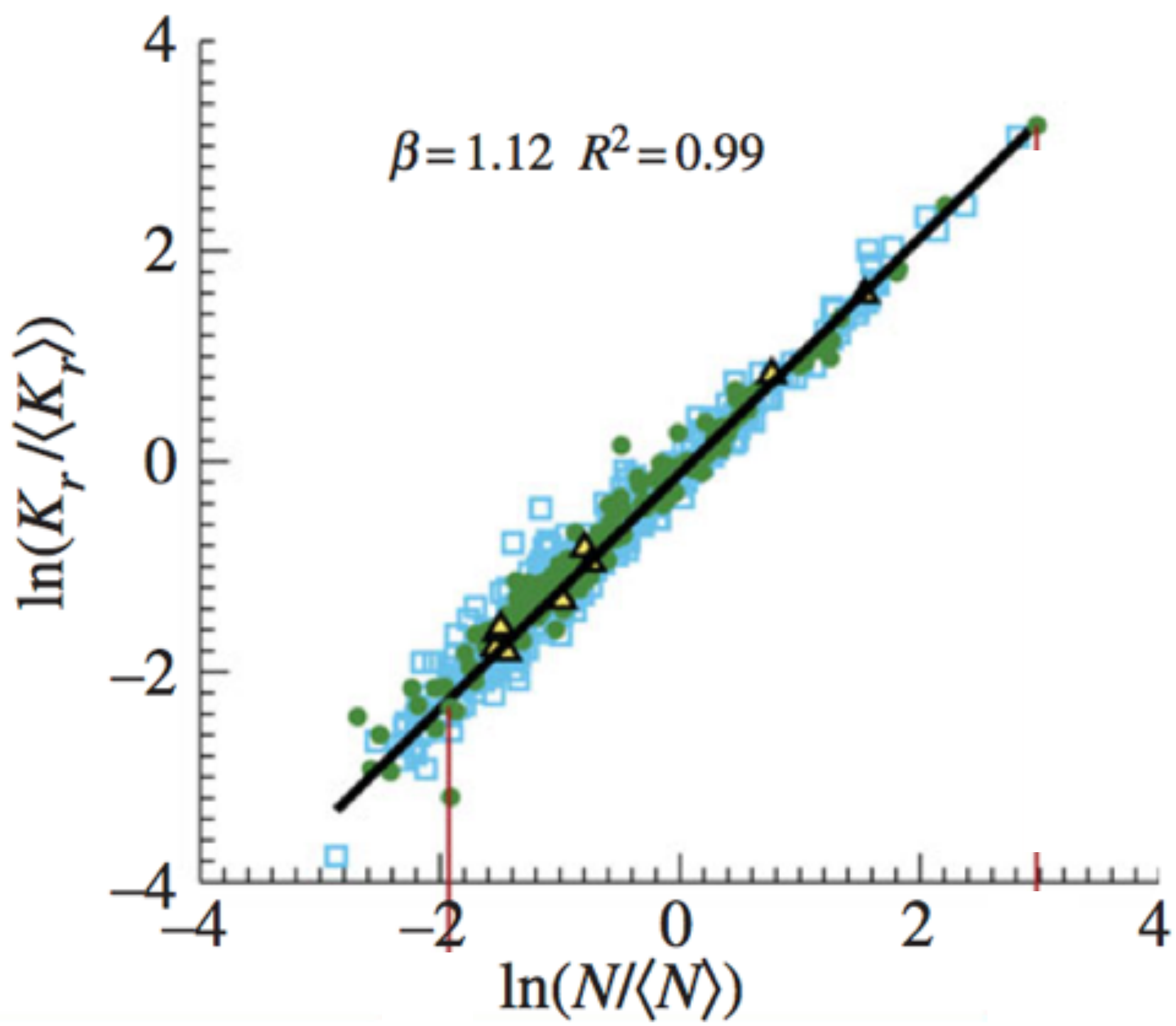
A: ratio of contacts between cities of size N_1 and N_2

e.g. $N_1/N_2 = 10^7$ and $A = 10$

$$\Rightarrow \quad \beta = 1.14$$

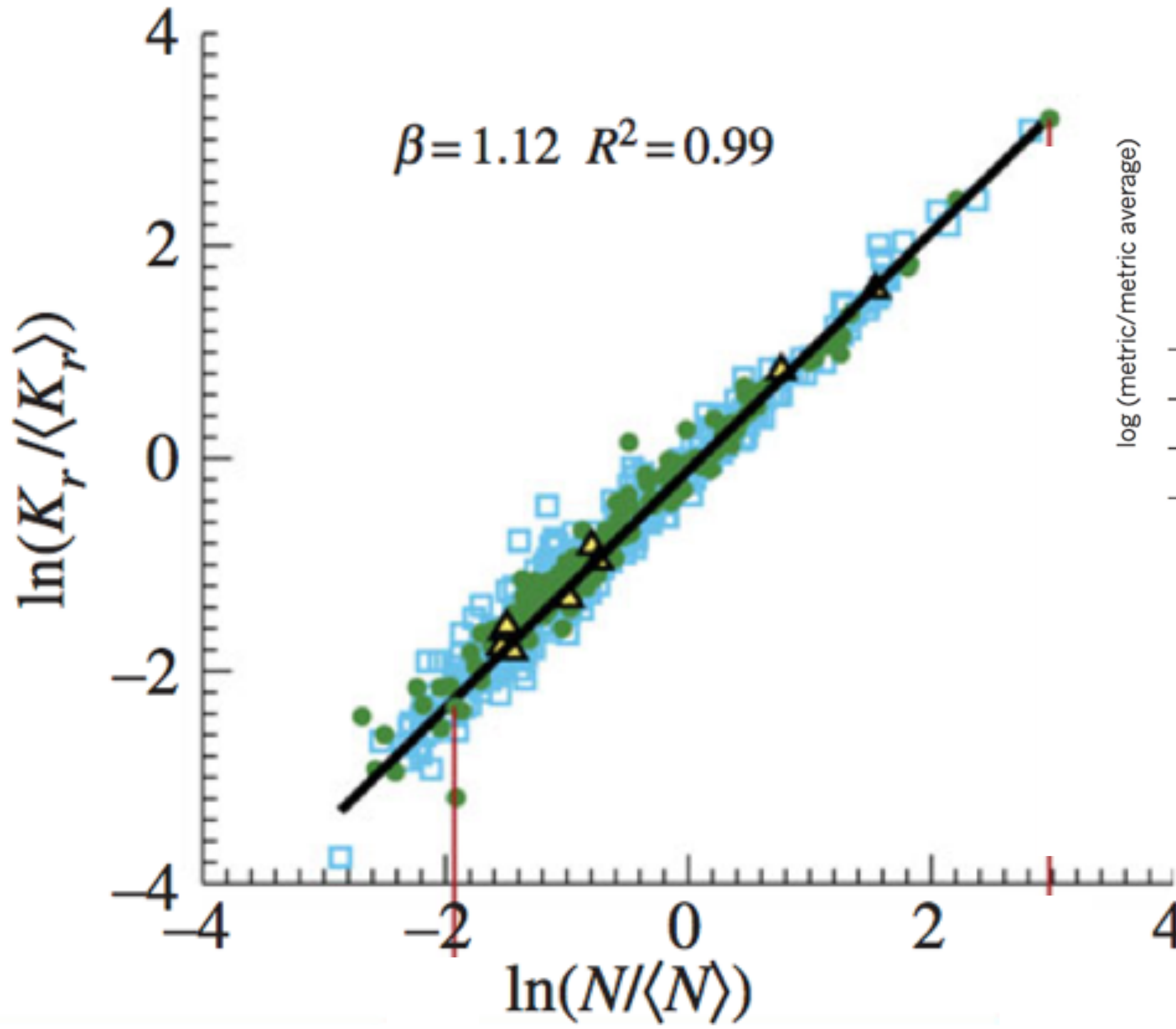
Scaling of human interactions

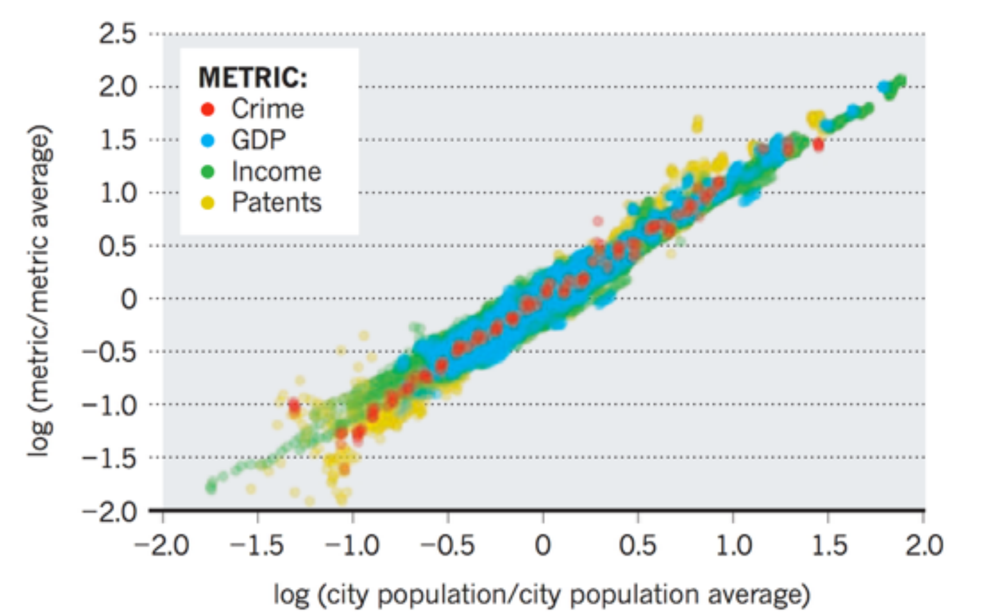
Cumulative degree



Scaling of human interactions

Cumulative degree





city definition	number	network type	ΔT (days)	Y	β	95% CI
Portugal						
statistical city	140	reciprocal	409	degree (K_r)	1.12	[1.11, 1.14]
				call volume (V_r)	1.11	[1.09, 1.12]
				number of calls (W_r)	1.10	[1.09, 1.11]
			92	degree (K_r)	1.10	[1.09, 1.11]
				call volume (V_r)	1.10	[1.08, 1.11]
				number of calls (W_r)	1.08	[1.07, 1.10]
		non-reciprocal	409	degree (K_r)	1.24	[1.22, 1.25]
				call volume (V_r)	1.14	[1.12, 1.15]
				number of calls (W_r)	1.13	[1.12, 1.14]
larger urban zone	9(8)	reciprocal	409	degree (K_r)	1.05	[1.00, 1.11]
				call volume (V_r)	1.11	[1.02, 1.20]
				number of calls (W_r)	1.10	[1.05, 1.15]
		non-reciprocal	409	degree (K_r)	1.13	[1.08, 1.18]
				call volume (V_r)	1.14	[1.05, 1.23]
				number of calls (W_r)	1.13	[1.08, 1.18]
municipality	293	reciprocal	409	degree (K_r)	1.13	[1.11, 1.14]
				call volume (V_r)	1.15	[1.13, 1.17]
				number of calls (W_r)	1.13	[1.11, 1.14]
UK						
urban audit city	24	reciprocal	31	degree (K)	1.08	[1.05, 1.12]
				degree, land-mobile (K_{lm})	1.14	[1.11, 1.17]
				call volume (V)	1.10	[1.07, 1.14]
				number of calls (W)	1.08	[1.05, 1.11]

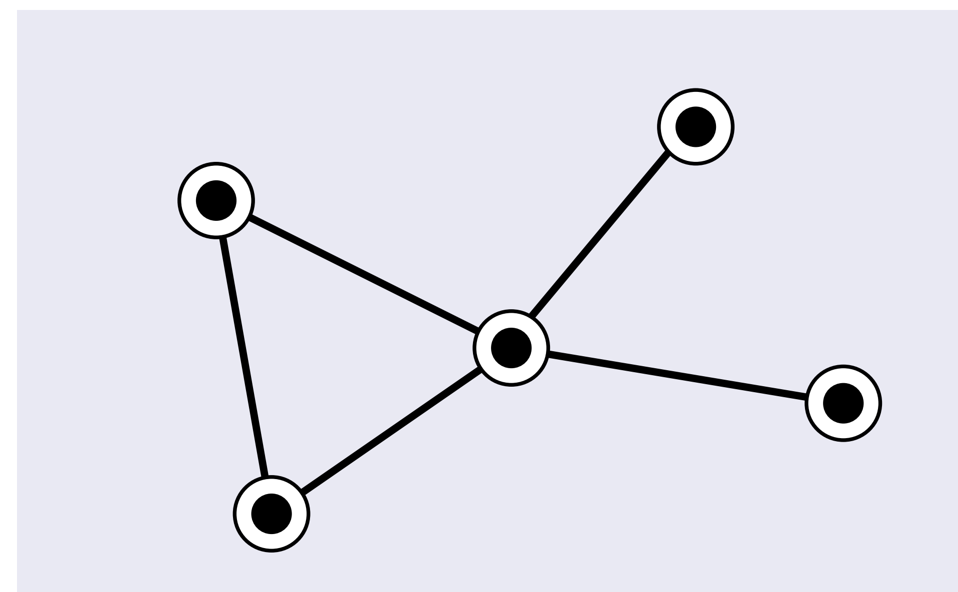
city definition	number	network type	ΔT (days)	Y	β	95% CI
Portugal						
statistical city	140	reciprocal	409	degree (K_r)	1.12	[1.11, 1.14]
				call volume (V_r)	1.11	[1.09, 1.12]
				number of calls (W_r)	1.10	[1.09, 1.11]
			92	degree (K_r)	1.10	[1.09, 1.11]
				call volume (V_r)	1.10	[1.08, 1.11]
				number of calls (W_r)	1.08	[1.07, 1.10]
		non-reciprocal	409	degree (K_r)	1.24	[1.22, 1.25]
				call volume (V_r)	1.14	[1.12, 1.15]
				number of calls (W_r)	1.13	[1.12, 1.14]
larger urban zone	9(8)	reciprocal	409	degree (K_r)	1.05	[1.00, 1.11]
				[illegible]	1.11	[1.02, 1.20]
				[illegible]	1.10	[1.05, 1.15]
		non-reciprocal	[illegible]	[illegible]	1.13	[1.08, 1.18]
				[illegible]	1.14	[1.05, 1.23]
				[illegible]	1.13	[1.08, 1.18]
municipality	293	reciprocal	[illegible]	[illegible]	1.13	[1.11, 1.14]
				[illegible]	1.15	[1.13, 1.17]
				number of calls (W_r)	1.13	[1.11, 1.14]
UK						
urban audit city	24	reciprocal	31	degree (K)	1.08	[1.05, 1.12]
				degree, land-mobile (K_{lm})	1.14	[1.11, 1.17]
				call volume (V)	1.10	[1.07, 1.14]
				number of calls (W)	1.08	[1.05, 1.11]

Nodal clustering

Clustering coefficient:
Probability that one‘s contacts are also connected with each other.

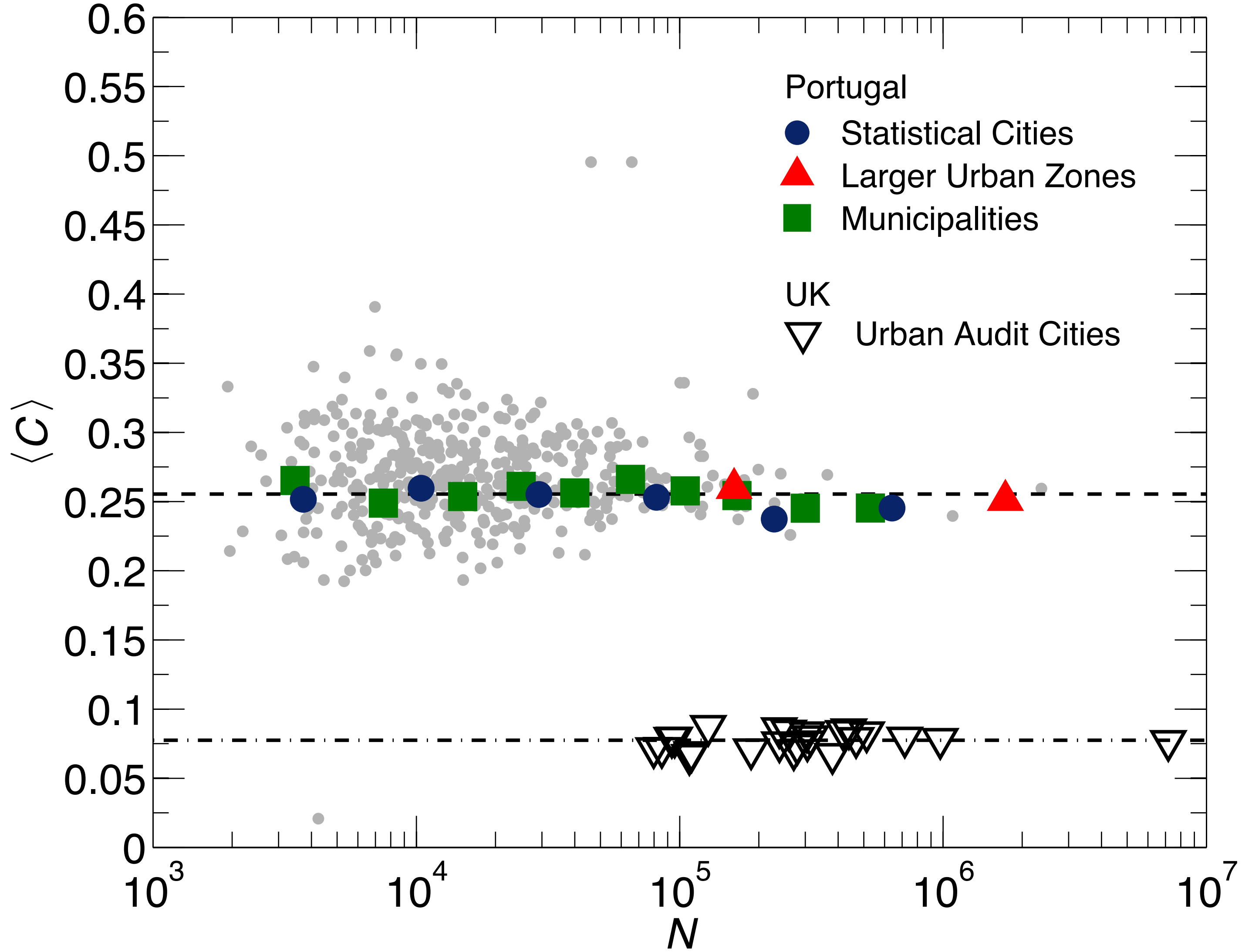
$$C_i \equiv 2z_i/\left[k_i(k_i-1)\right]$$

z_i Number of links between the k_i neighbours
k_i Degree of node i



As larger cities provide a larger pool of people, the clustering coefficient should decrease if contacts were established at random.

Nodal clustering



- Average clustering is an invariant of city size.
- Even in large cities we live in groups that are as tightly knit as those in small towns or 'villages'.

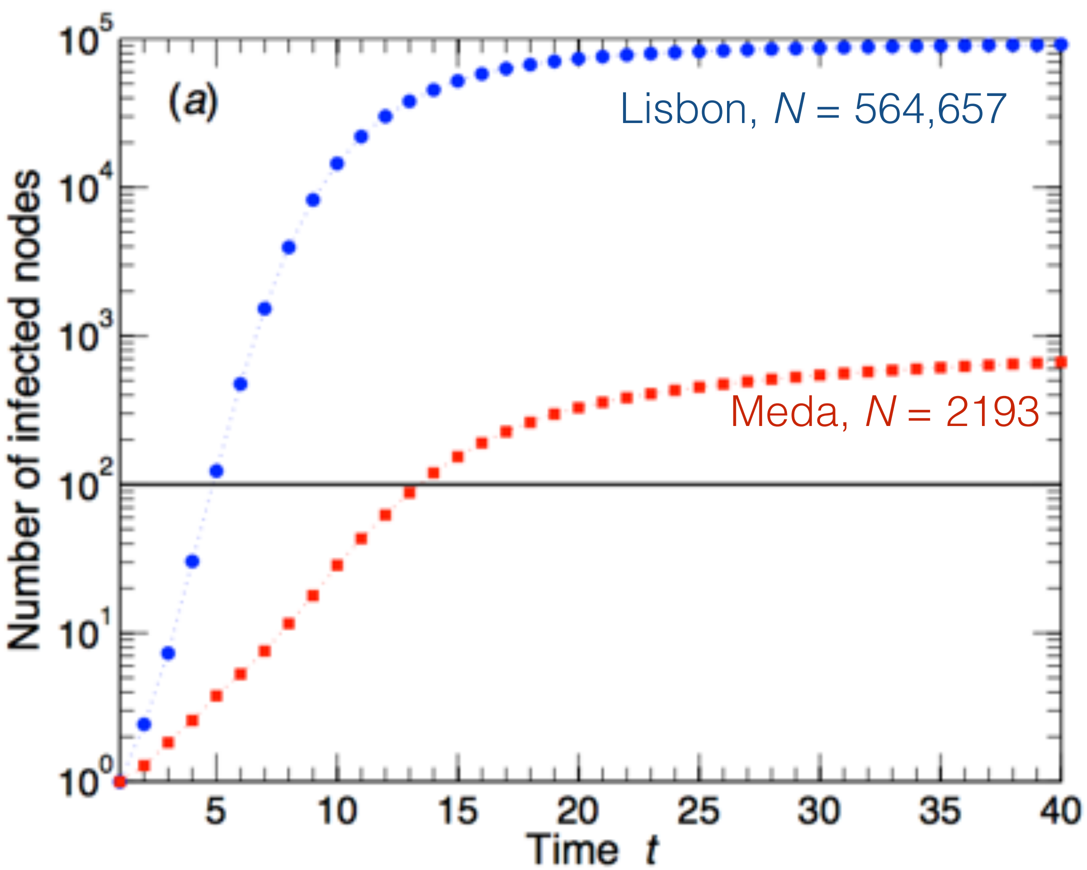
Schläpfer M. et al., J Royal Soc Interface, 11(98):20130789, 2014

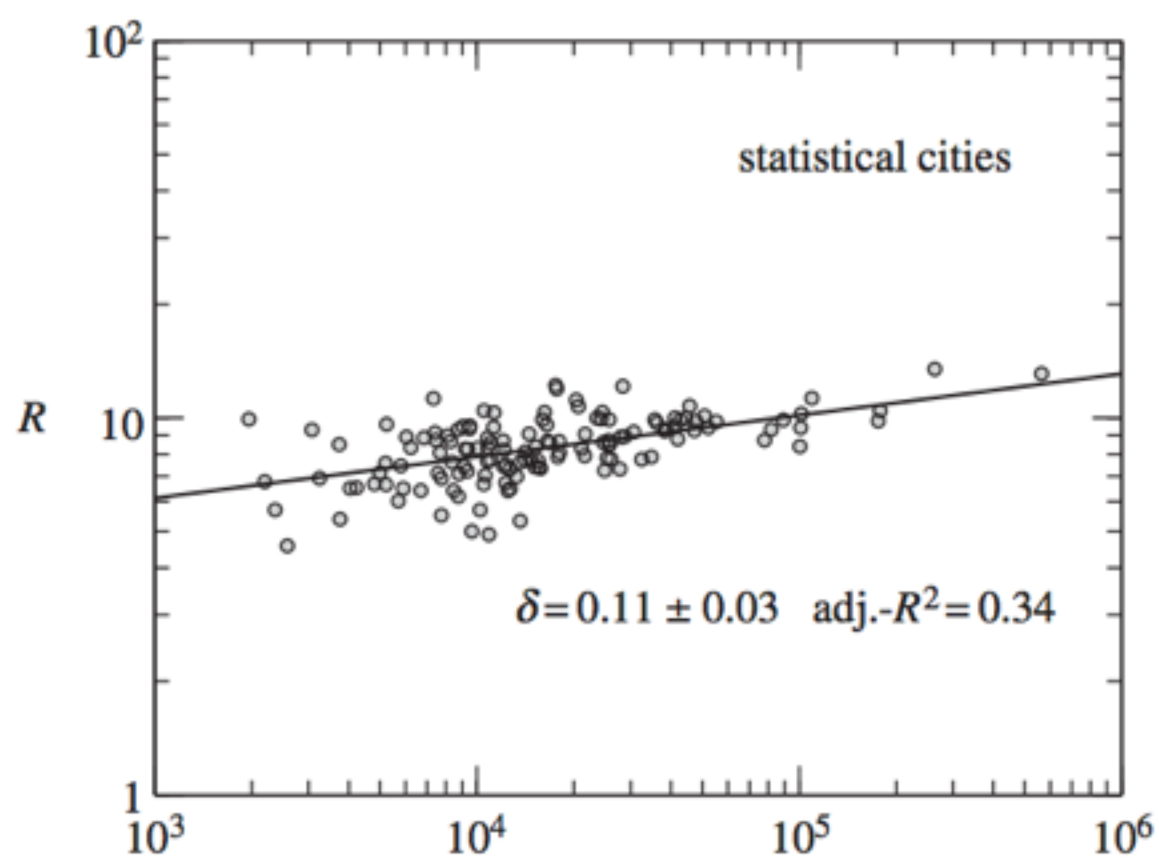
Acceleration of spreading processes

Susceptible-infected (SI) model

$$P_{ij} \propto v_{ij}$$

P_{ij} Transmission probability
v_{ij} Call volume between user i and user j





Potential ‚hidden‘ biases

Test on different data sets

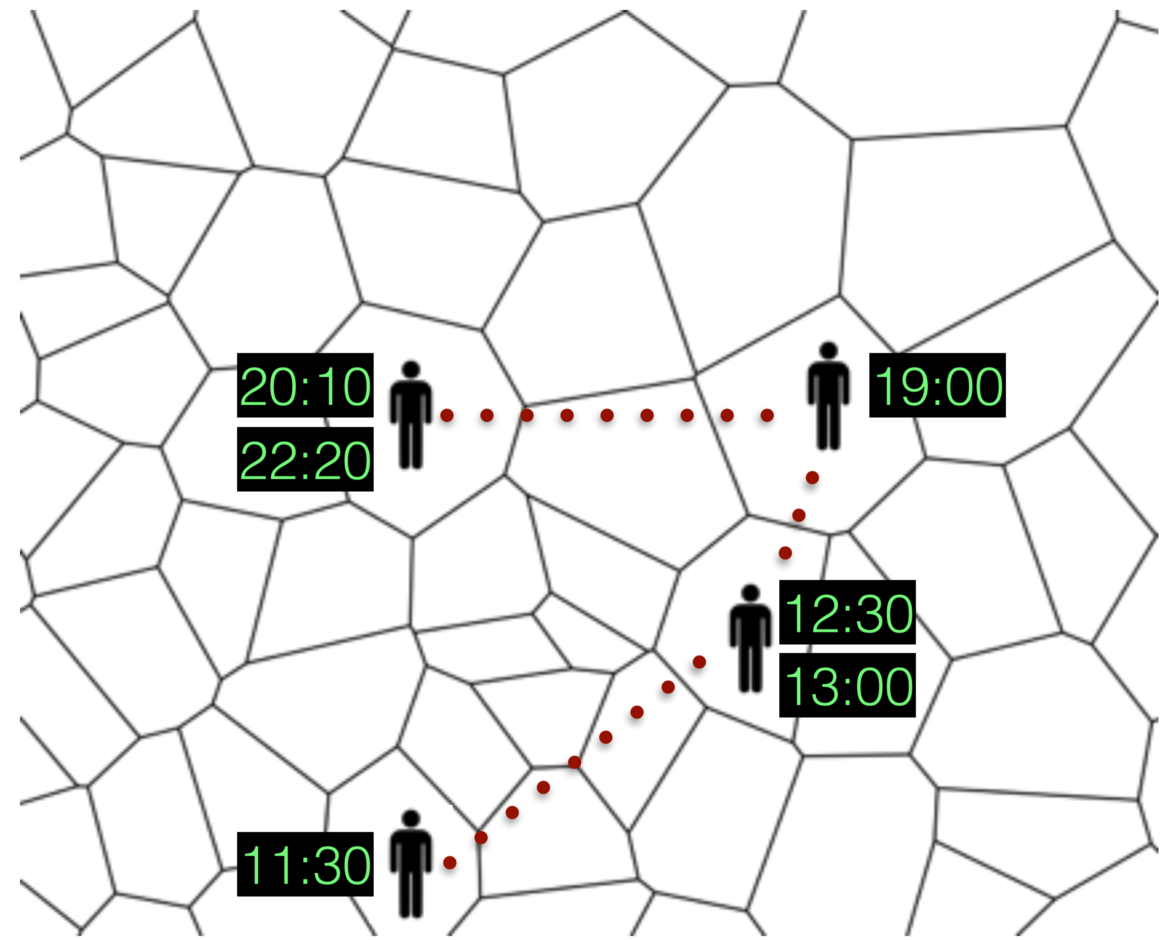
- UK, mobile phones and landlines (Schläpfer et al. 2014)
- Ivory Coast, mobile phones (Andris and Bettencourt, 2014)
- „Unnamed“ European Country, mobile phones (Llorente, 2015)
- US and Europe, Twitter data (Tizzoni, 2015)

Part II: Movement of people in cities

Individual trajectories

User ID, Timestamp, Cell tower ID

```
1,2013-01-24 11:30:00,599
1,2013-01-24 12:30:00,608
1,2013-01-24 13:00:00,608
1,2013-01-24 19:00:00,446
1,2013-01-24 20:10:00,323
1,2013-01-24 20:30:00,323
1,2013-01-24 22:00:00,323
1,2013-01-24 22:20:00,323
```



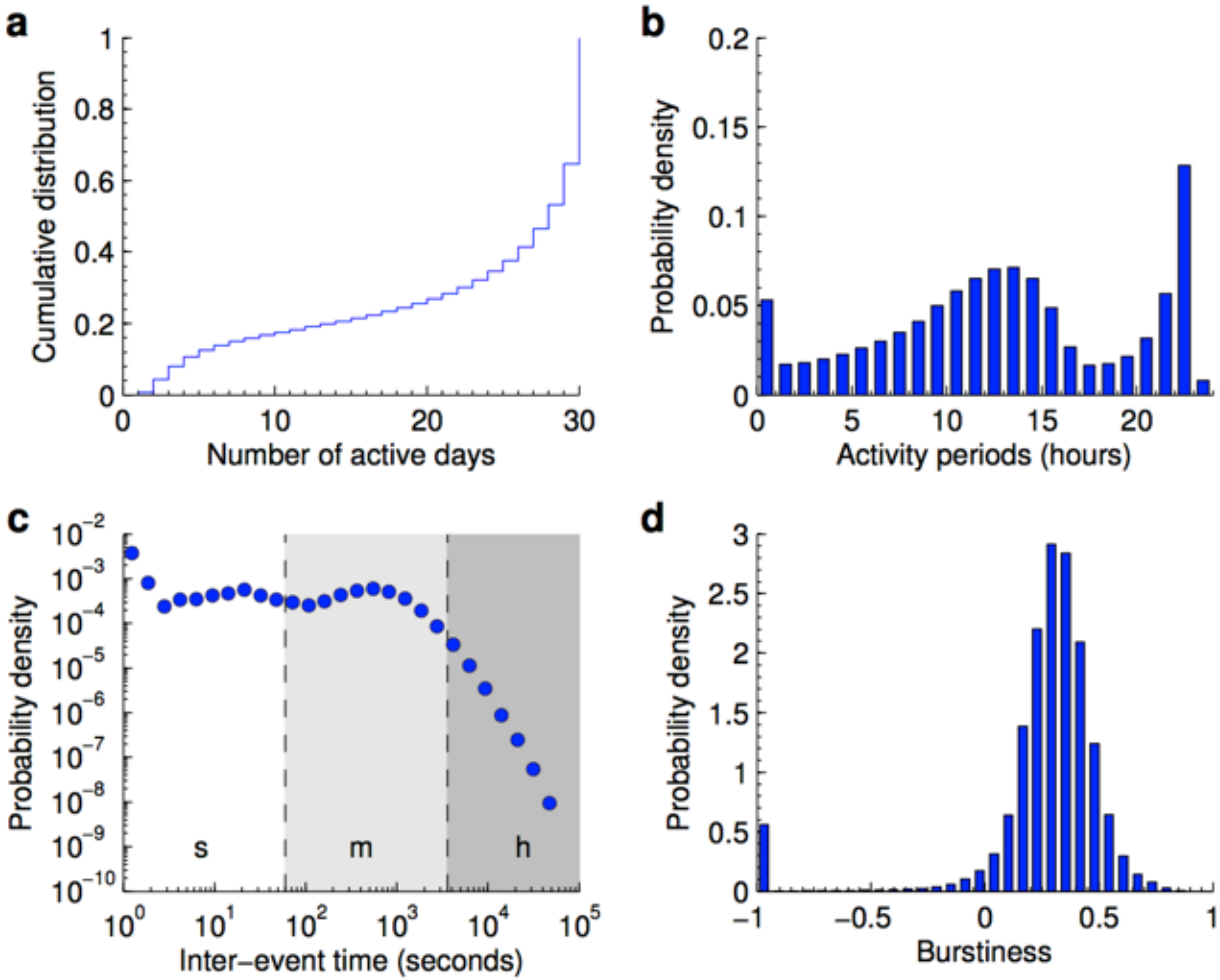
„*Collective*“ movements

- *How many* people visit a given location?
- From *how far* do they come?
- *How often* do they visit?

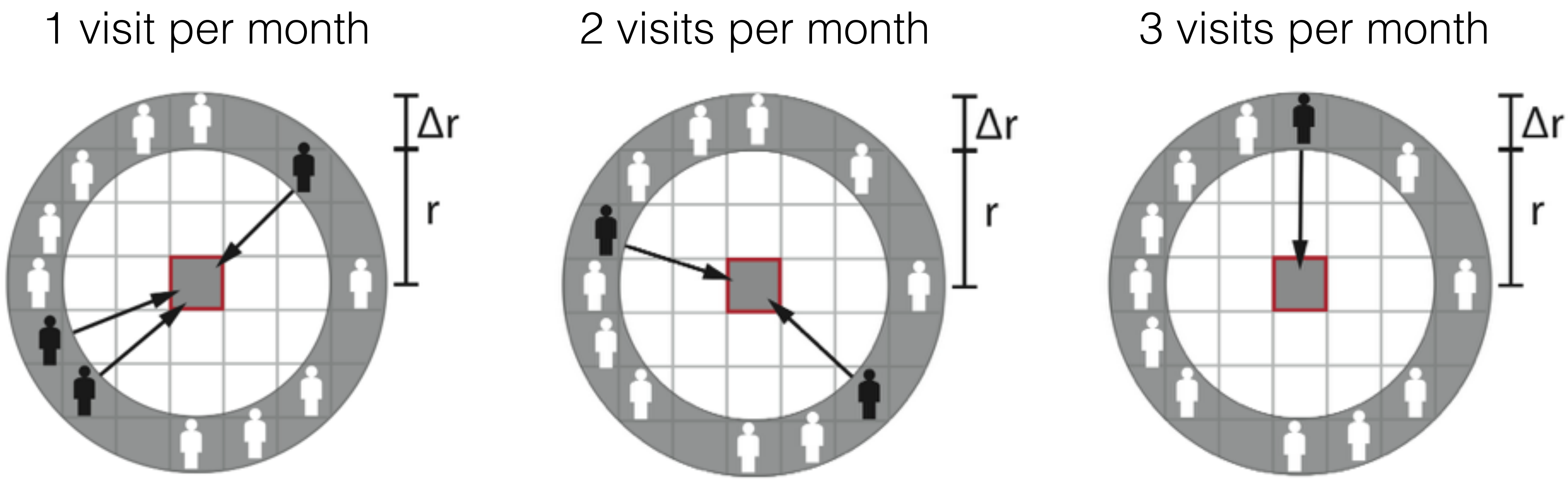
Lets look into the data!

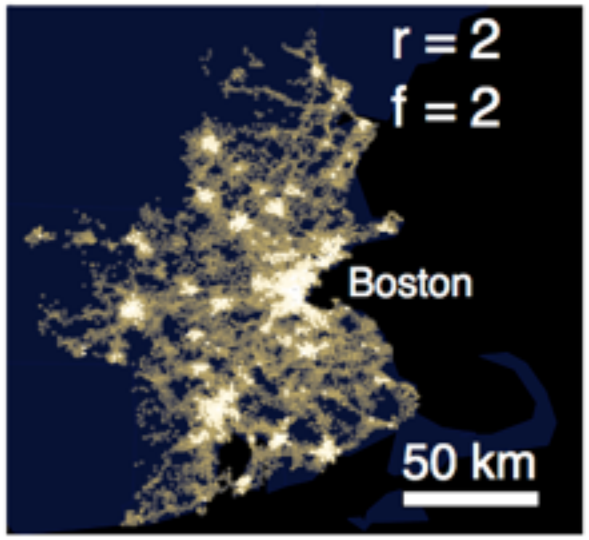
- Greater Boston area
- 2 Mio. mobile phone users over 4 months
- 10^8 location based records per month (triangulation)
- 14,450 locations (1km x 1km grid cells)

Sampling statistics



Quantifying the attractiveness of locations

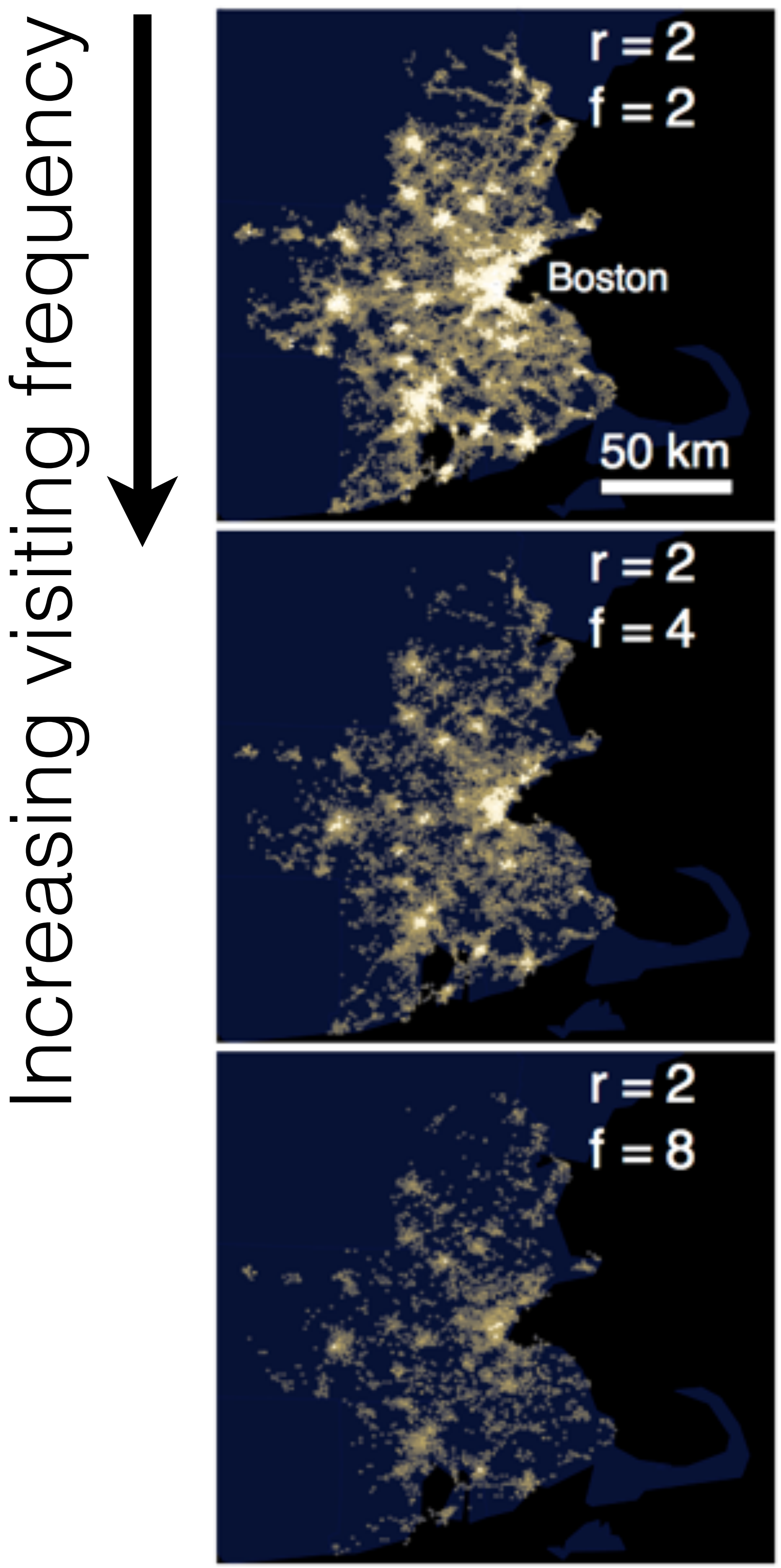


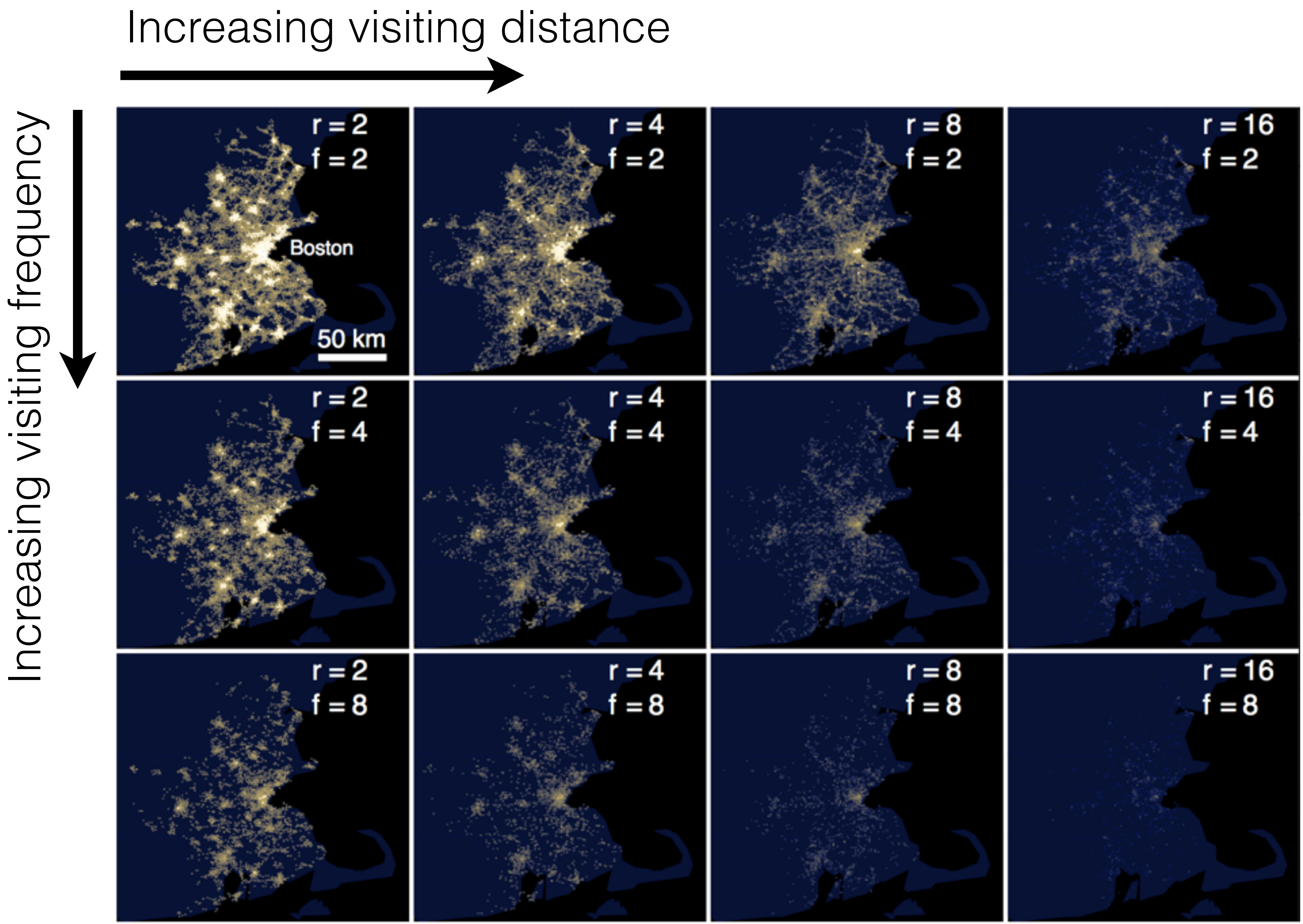


r visiting distance (km)
f visiting frequency (visits per month)

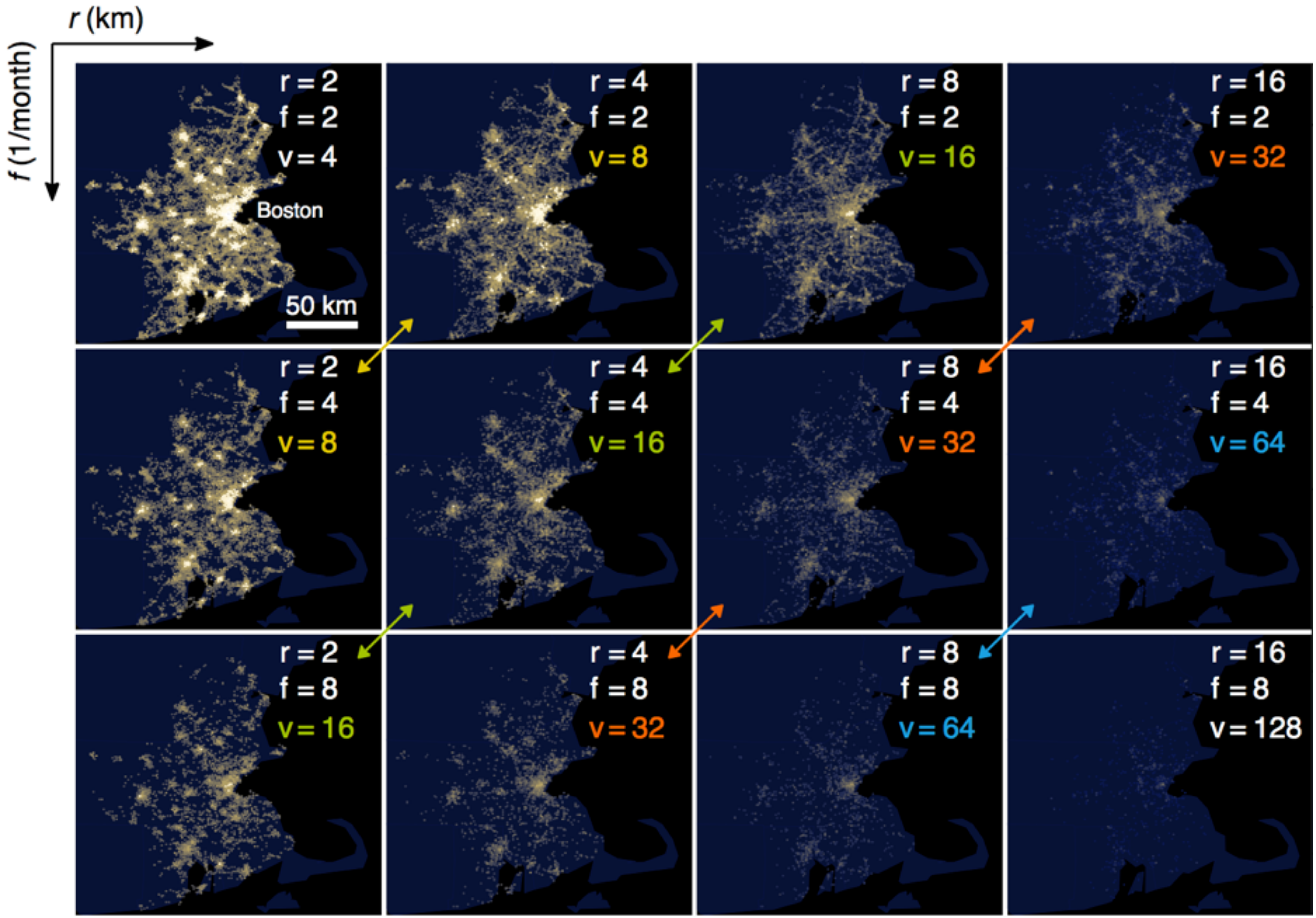
Brightness of pixel: number of visitors

Increasing visiting distance

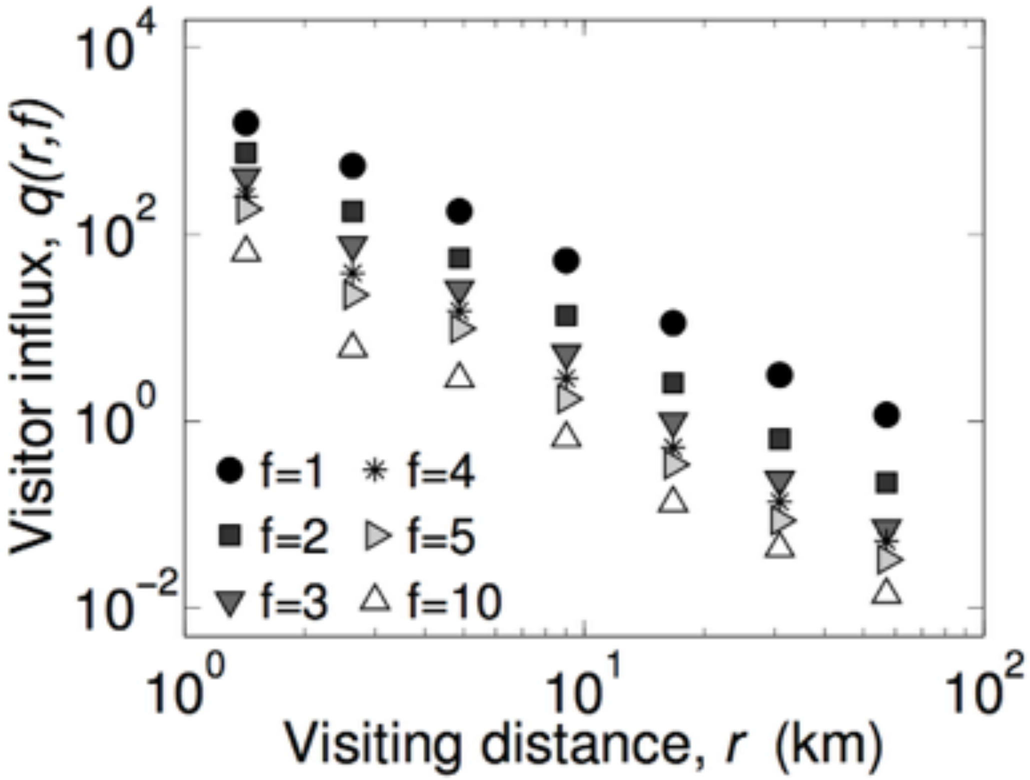

Increasing visiting frequency
r = 2
f = 2
Boston
50 km
r = 2
f = 4
r = 2
f = 8



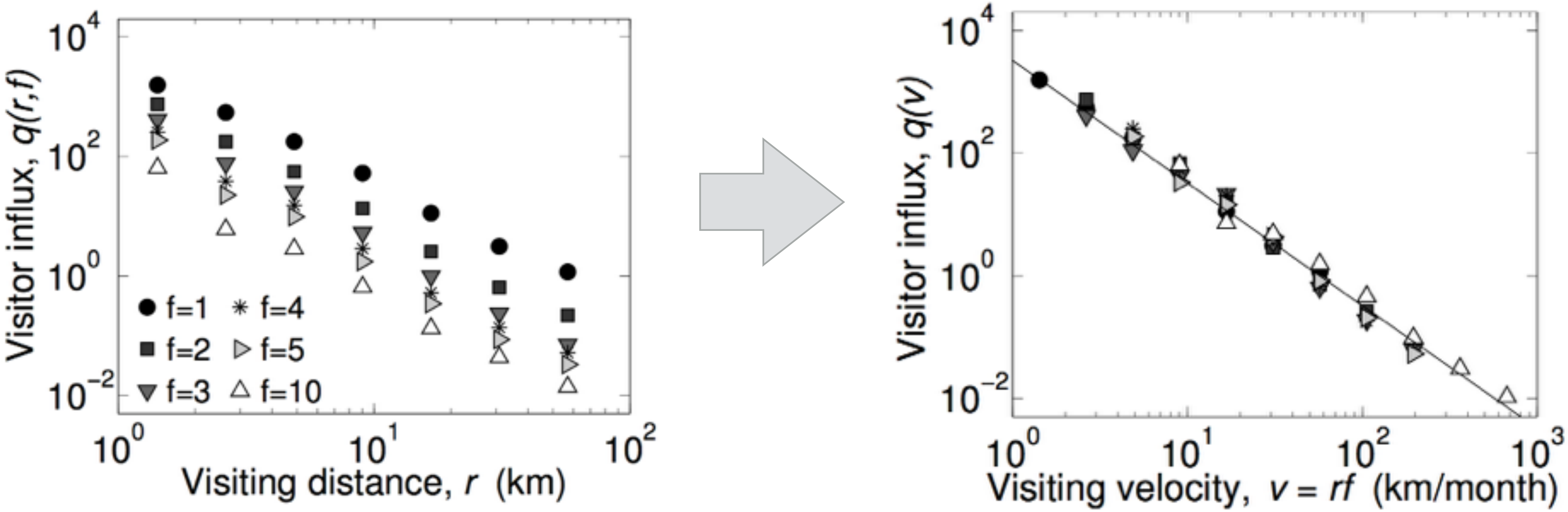
Increasing visiting distance
Increasing visiting frequency
r = 2
f = 2
Boston
50 km
r = 4
f = 2
r = 8
f = 2
r = 16
f = 2
r = 2
f = 4
r = 4
f = 4
r = 8
f = 4
r = 16
f = 4
r = 2
f = 8
r = 4
f = 8
r = 8
f = 8
r = 16
f = 8



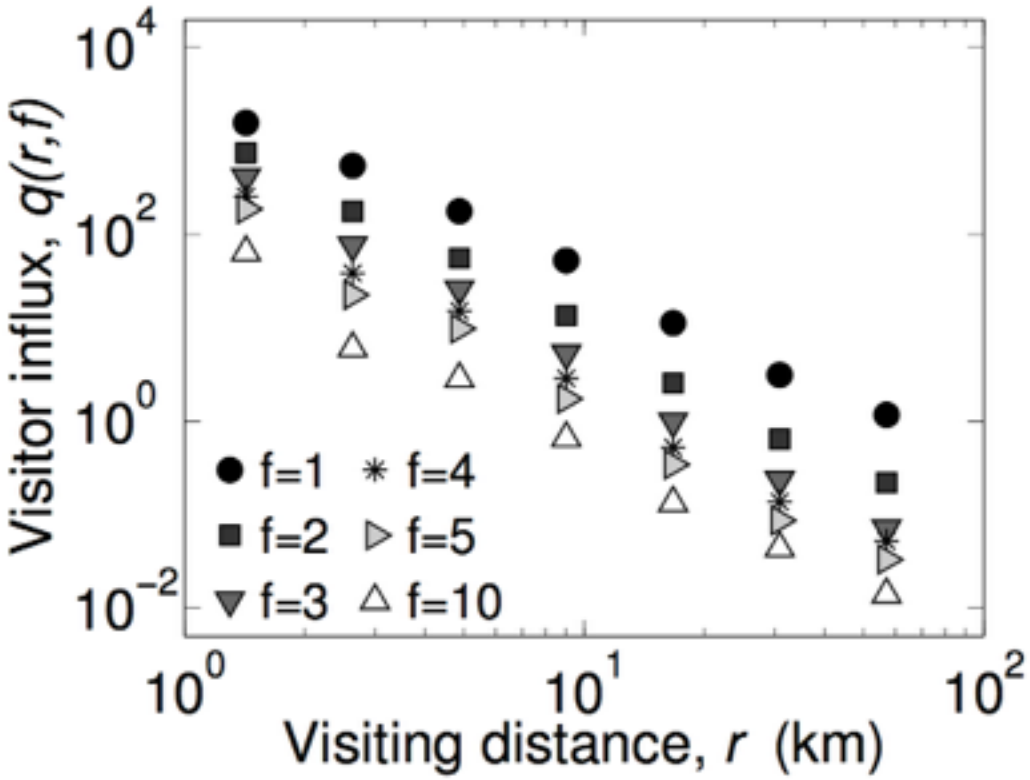
r (km)
f (1/month)
r = 2
f = 2
v = 4
Boston
50 km
r = 4
f = 2
v = 8
r = 8
f = 2
v = 16
r = 16
f = 2
v = 32
r = 2
f = 4
v = 8
r = 4
f = 4
v = 16
r = 8
f = 4
v = 32
r = 16
f = 4
v = 64
r = 2
f = 8
v = 16
r = 4
f = 8
v = 32
r = 8
f = 8
v = 64
r = 16
f = 8
v = 128

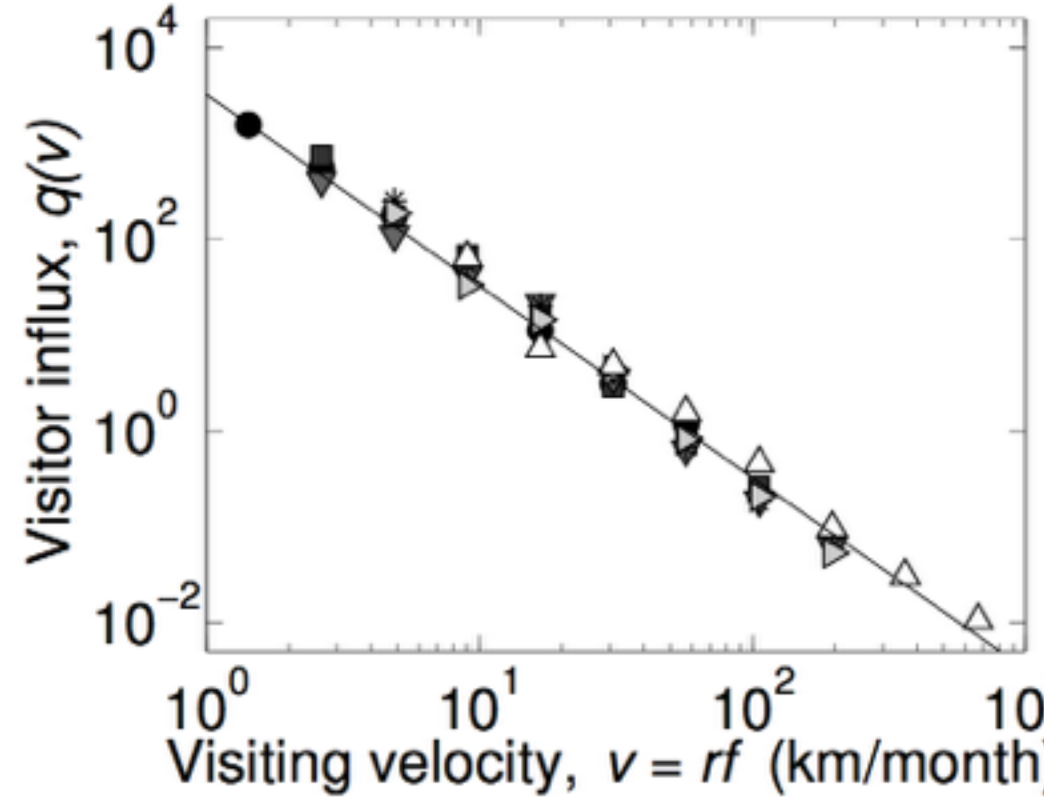


Visitor influx, q(r,f)
Visiting distance, r (km)
f=1
f=2
f=3
f=4
f=5
f=10



Visitor influx, q(r,f)
f=1
f=2
f=3
f=4
f=5
f=10
Visiting distance, r (km)
Visitor influx, q(v)
Visiting velocity, v = rf (km/month)





Example (v = 20km/month):

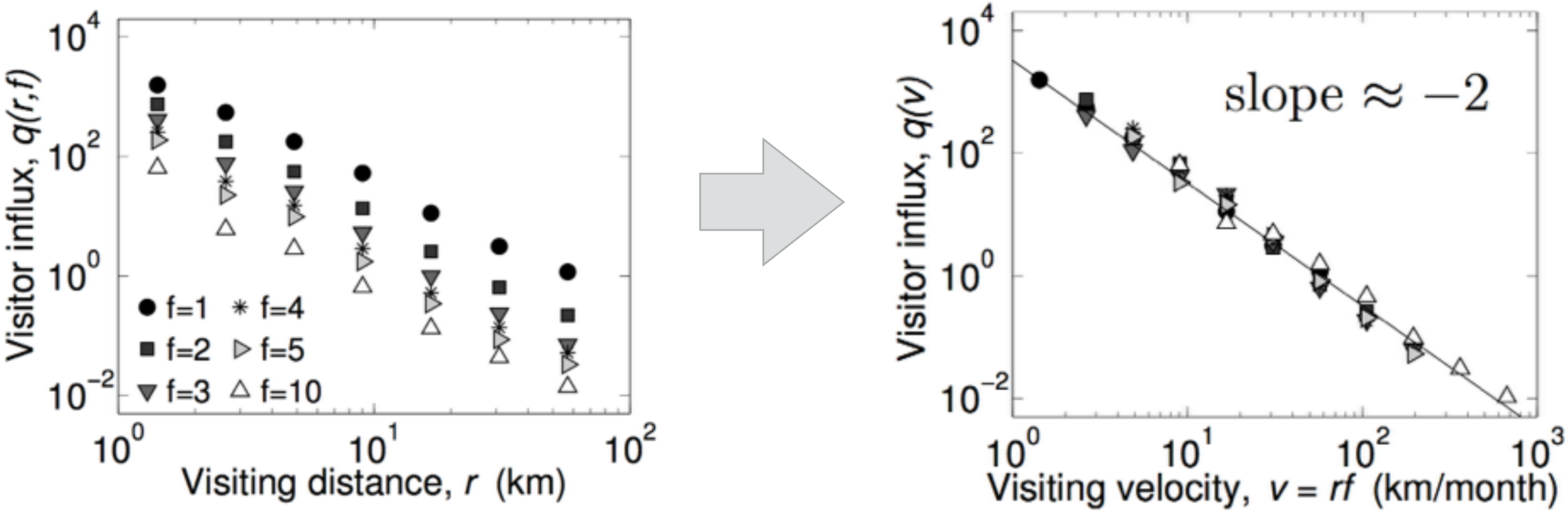
number of visitors coming from 5 km and 4 times a month

=

number of visitors coming from 10 km and 2 times a month

=

number of visitors coming from 20 km and once a month



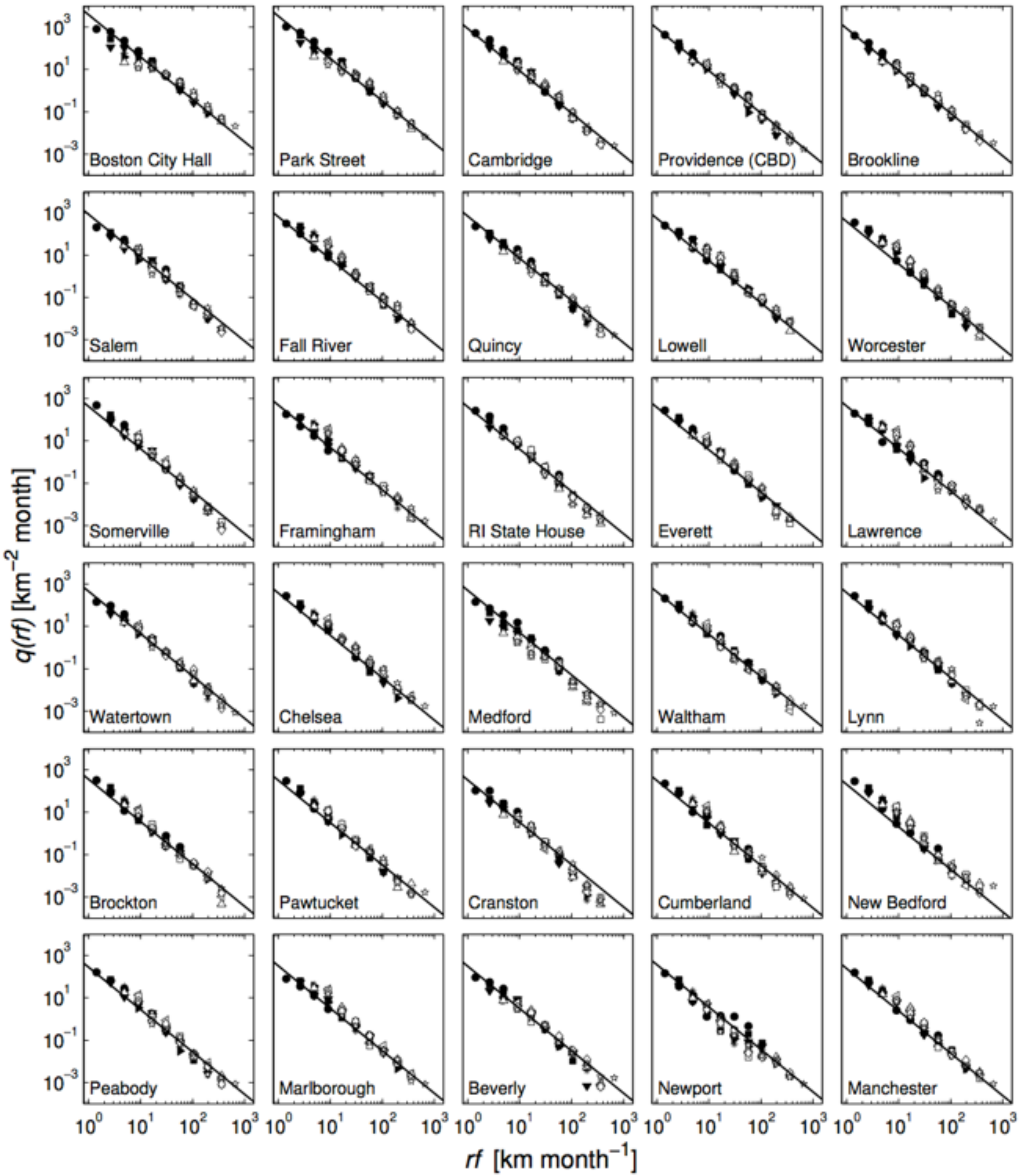
$$\text{People influx} \propto v^{-2}$$

Can be derived from basic principles/assumptions:

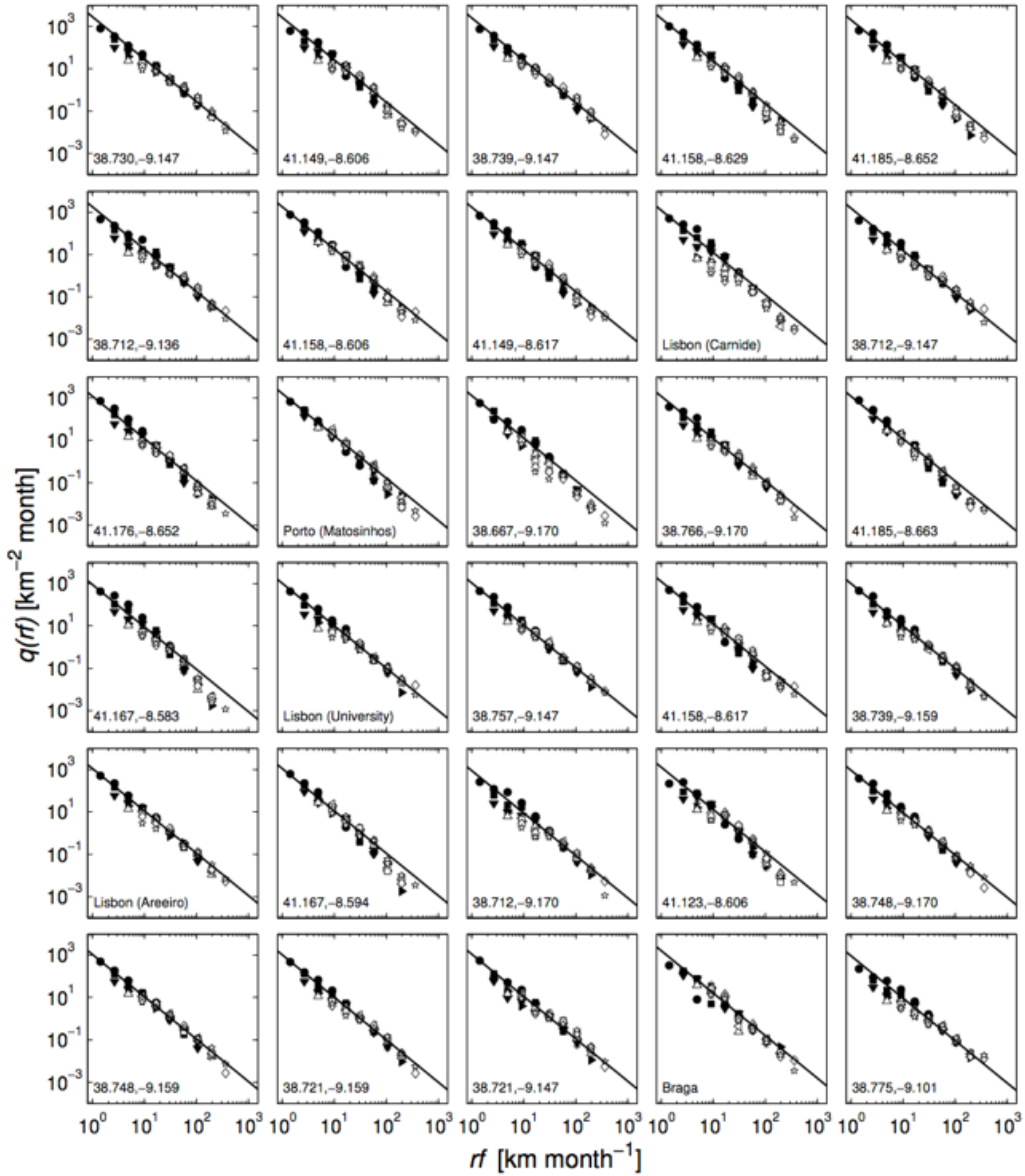
People, on average, take shortest route to locations

The distance-frequency distribution can be obtained by just counting the total rate at which people visit the center.

Greater Boston

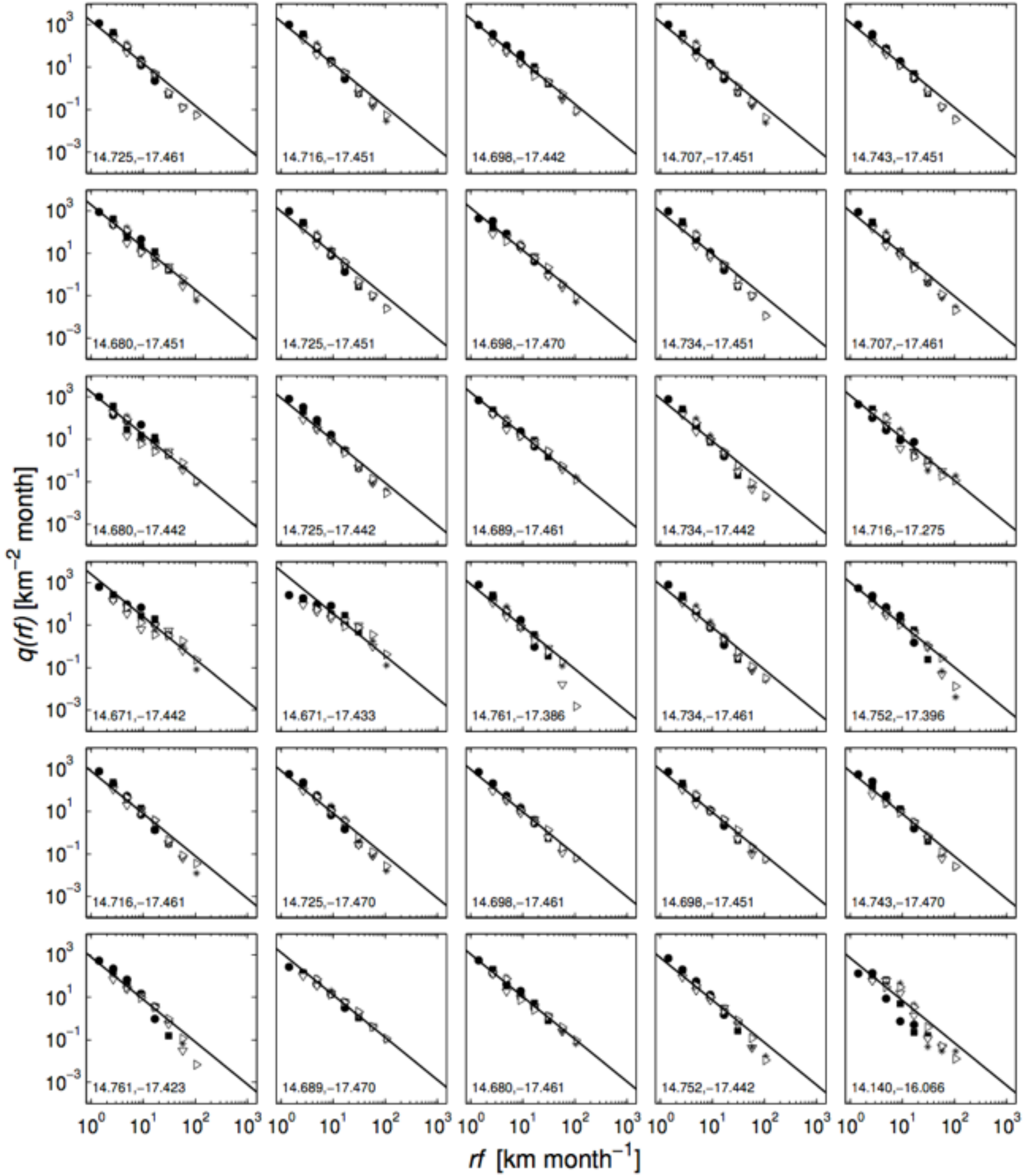


Portugal



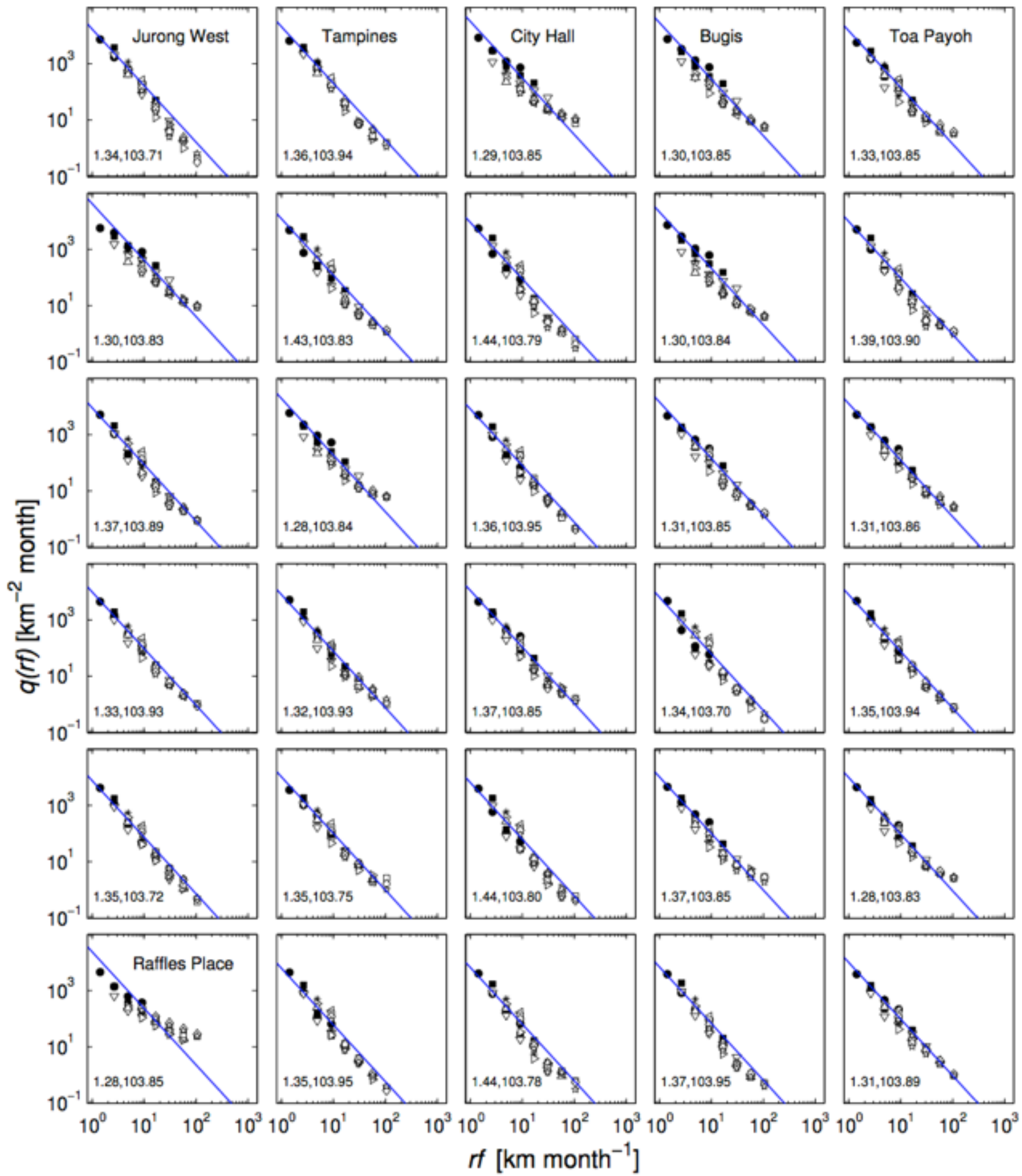
38.730,−9.147
41.149,−8.606
38.739,−9.147
41.158,−8.629
41.185,−8.652
38.712,−9.136
41.158,−8.606
41.149,−8.617
Lisbon (Carnide)
38.712,−9.147
41.176,−8.652
Porto (Matosinhos)
38.667,−9.170
38.766,−9.170
41.185,−8.663
41.167,−8.583
Lisbon (University)
38.757,−9.147
41.158,−8.617
38.739,−9.159
Lisbon (Areeiro)
41.167,−8.594
38.712,−9.170
41.123,−8.606
38.748,−9.170
38.748,−9.159
38.721,−9.159
38.721,−9.147
Braga
38.775,−9.101
q(rf) [km−2 month]
rf [km month−1]

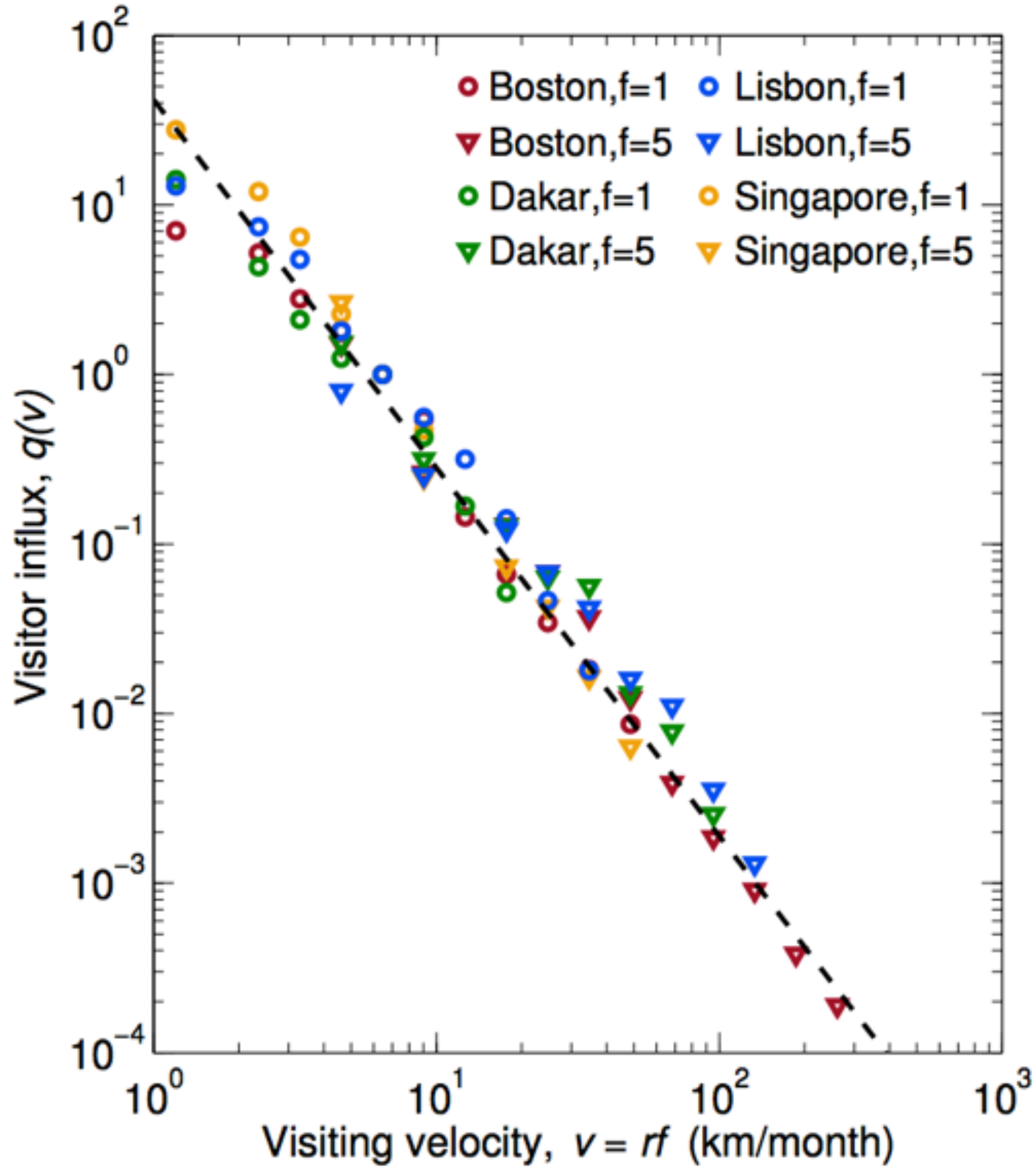
Senegal



14.725,-17.461
14.716,-17.451
14.698,-17.442
14.707,-17.451
14.743,-17.451
14.680,-17.451
14.725,-17.451
14.698,-17.470
14.734,-17.451
14.707,-17.461
14.680,-17.442
14.725,-17.442
14.689,-17.461
14.734,-17.442
14.716,-17.275
14.671,-17.442
14.671,-17.433
14.761,-17.386
14.734,-17.461
14.752,-17.396
14.716,-17.461
14.725,-17.470
14.698,-17.461
14.698,-17.451
14.743,-17.470
14.761,-17.423
14.689,-17.470
14.680,-17.461
14.752,-17.442
14.140,-16.066
q(rf) [km^{-2} month]
rf [km month^{-1}]

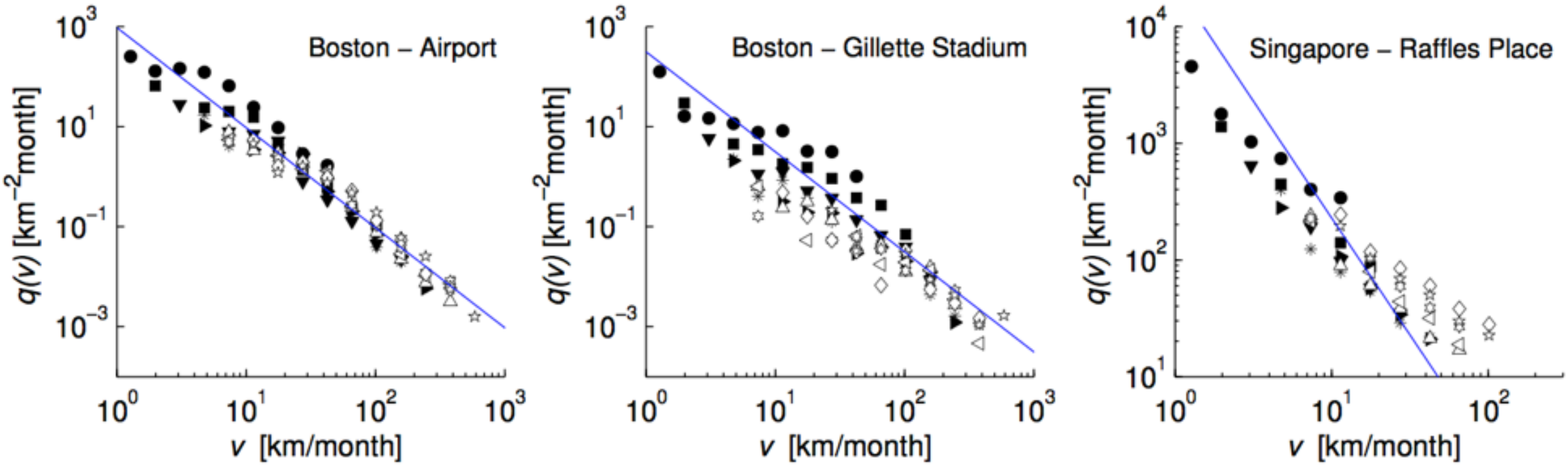
Singapore

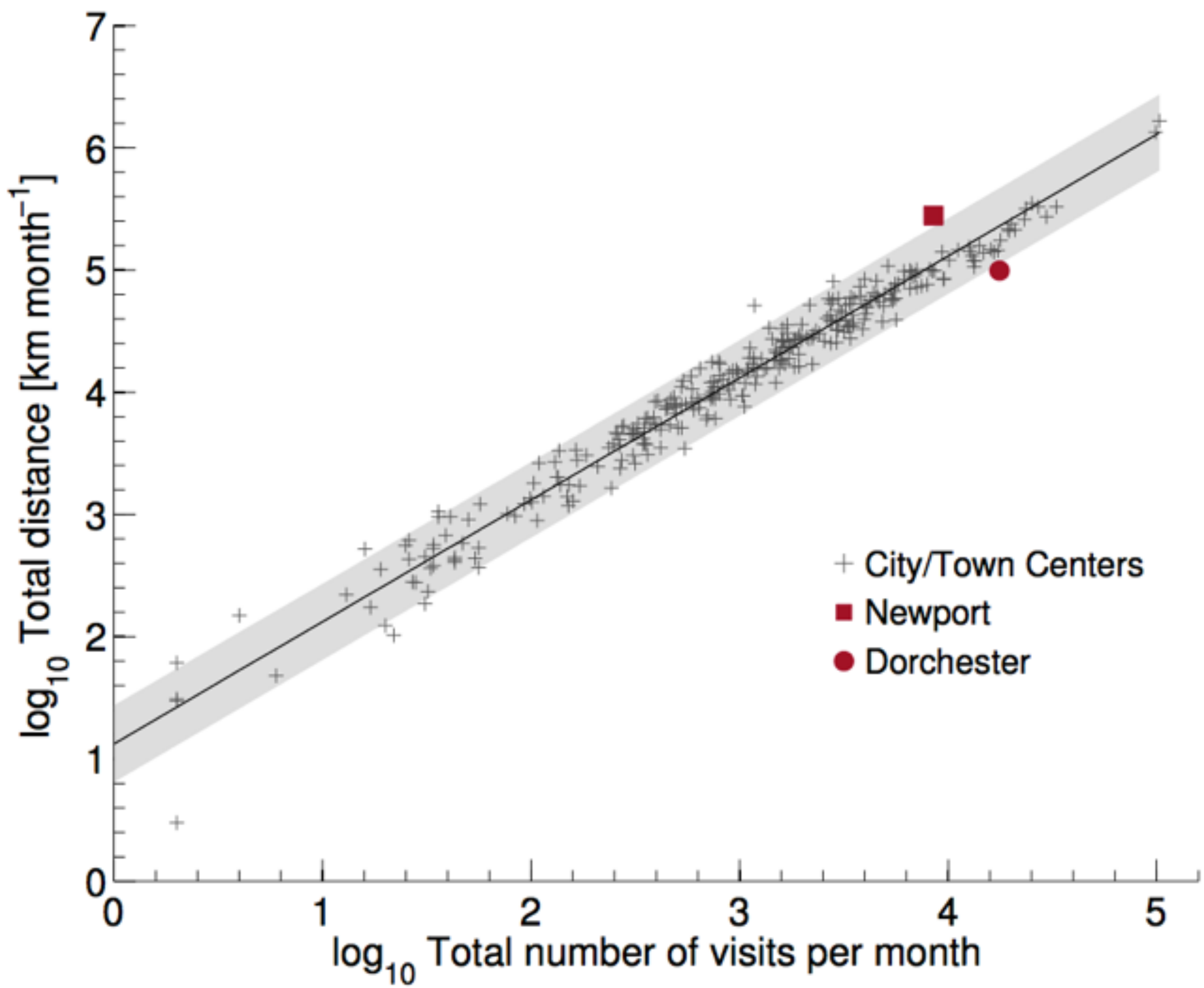

Jurong West
Tampines
City Hall
Bugis
Toa Payoh
Raffles Place
1.34,103.71
1.36,103.94
1.29,103.85
1.30,103.85
1.33,103.85
1.30,103.83
1.43,103.83
1.44,103.79
1.30,103.84
1.39,103.90
1.37,103.89
1.28,103.84
1.36,103.95
1.31,103.85
1.31,103.86
1.33,103.93
1.32,103.93
1.37,103.85
1.34,103.70
1.35,103.94
1.35,103.72
1.35,103.75
1.44,103.80
1.37,103.85
1.28,103.83
1.28,103.85
1.35,103.95
1.44,103.78
1.37,103.95
1.31,103.89
q(rf) [km^-2 month]
rf [km month^-1]



Boston,f=1
Boston,f=5
Dakar,f=1
Dakar,f=5
Lisbon,f=1
Lisbon,f=5
Singapore,f=1
Singapore,f=5
Visitor influx, q(v)
Visiting velocity, v = rf (km/month)

Locations with ‚anomalous‘ behavior





log10 Total distance [km month-1]
log10 Total number of visits per month
City/Town Centers
Newport
Dorchester
0
1
2
3
4
5
6
7

Take home:
urban ‚big‘ data

„Don’t“s	„Do“s
give new insights per se	
are free of ‚hidden‘ biases	

Take home:
urban ‚big' data

„Don't"s	„Do"s
give new insights per se	allow testing „old" ideas - reveal hidden regularities
are free of ‚hidden' biases	cover large parts of the population
	objective measurements

In collaboration with..

Luis Bettencourt, SFI

Geoffrey West, SFI

Carlo Ratti, MIT

Sebastian Grauwin, MIT

Michael Szell, NEU

Mathias Raschke, Raschke Engineering

Zbigniew Smoreda, Orange Telecommunication

Thanks!

Markus Schläpfer

web.mit.edu/schlmark/www/