CITATION:

Pallares-Barbera, Montserrat (2017). Big Data and Geography. Spatial analysis from the demand side. Big Data y Geografía. Análisis Espacial desde la Demanda. Conference at the XXV Congreso Asociación de Geógrafos Españoles. Organizing Group: Tecnologías de la Información Geográfica, AGE. Mesa Temática: Big data: nuevas fuentes de información en la investigación geográfica. Madrid: October, 27, 9.00-11.00.

Big Data y Geografía. Análisis Espacial desde la Demanda

Montserrat Pallares-Barbera

Universitat Autònoma de Barcelona
Departamento de Geografía

Mesa Temática: Big data: nuevas fuentes de información en la investigación geográfica

XXV Congreso Asociación de Geógrafos Españoles

Grupo organizador: Tecnologías de la Información Geográfica

Madrid, 27 octubre 2017, 9-11

Sumario

- Big data y Geografía, Análisis Espacial
- Hipótesis
- Modelización
- Datos
- Metodología
- Problemas que aparecen con Big data
- Problemas que resuelve Big data
- Patrones geográficos

Los fundamentos de la geografía

- Patrones espaciales (Hartshorne (1939), Harvey (1963), Friedman and Alonso (1964), Berry (1973), Chorley and Haggett (1965)
- Análisis especial
- Regularidad y organización de personas y actividades (patrones)

Cambios en el comportamiento de la población

- Producción de gran cantidad de datos desde el propio usuario (User Generated Data, UGD)
  (Snijders, Matzat, and Reips 2012; Targio et al. 2015)
  - Geolocalización de los datos (GPS)
  - Contenido semántico

- Usos de Tecnologías de la Información y de la Comunicación (TIC) sobretodo Smartphones

Tipos de datos

- Fuentes secundarias a fuentes primarias
- Big data y Scarce data (estadísticas producidas por instituciones)
- “can Big Data and data-driven methods lead to significant discoveries in geographic research?” (Miller and Goodchild (2015, p. 450)

Twitter

- Web-based social platform currently used by 284 million of monthly active users (Twitter 2015)
- It is used by only 21% of total population in USA (Greenwood, Perin and Duggan 2016)
- Time span/Geolocation

Case study forming

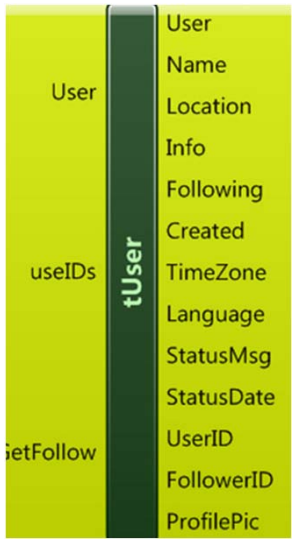
- **Objetivo**: In Barcelona, “how different types of users of public open spaces behave in their visiting of the city”
- **Hipótesis**: Each type of locals and visitors will choose to go to different location; depending to: its cultural baggage, country of origin, the day and hour of their visit. Each type will show specific spatial patterns and discover new social behaviors
- **Modelización**:
  - the differences between the number of tweets sent depending on the day of the week
  - the differences between the number of tweets sent depending on the hour of the day
- **Datos**: data from users of Twitter. Origin-Destination matrix

Caso de studio, metodología

- **Metodología**:
  - Hypothesis testing Chi-Squared test with FDR (false discovery rate) adjustment for multiple comparisons,
  - The statistical analysis was performed using R v3.2.4. For all statistical tests a nominal significance level of 5% ($p < 0.05$).
  - Data collection is based on the use of REST API (Twitter REST API, 2016),
  - Twitter Search Engine (TSE allows collection and storage of data for unlimited periods of time, Dinkic et al. 2016).

Twitter data content

ID	Message	User from	User to	Date	Lat	Lon
1	⛷❄▯ @ Outside - Calvin Harris http://t.co/14oG7mPxwE	albasanchoo95	fjdh	Wed Jan 07 08:55:20 +0000 2015	41.47267	2.27114
2	Nos aventuramos a las rebajas... Tengo miedo xD	zaidasphyxiated		Wed Jan 07 08:55:18 +0000 2015	41.52691	2.229489
3	#cyberparks #TU1306	e_frola		Fri Jan 09 10:29:08 +0000 2015	41.50982	2.229217
4	.....	.....	.....	.....	.....	.....
n	.....	.....	.....	.....	.....	.....



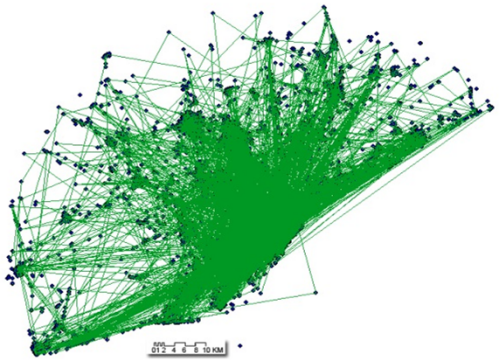
January 7-19, 2015. The total amount of records was 126,288 Tweets, of which 103,404 (81.88 %) were georeferenced.

January 13-20, 2017. 15-17000 tweets per day (TSE).

Results

Figure 2. Twitter messages. Barcelona metropolitan area (BMA) (a), and Barcelona city (BCN) (b)

a. O/D matrix BMA, January 7-19, 2015



b. January 13-20, 2017

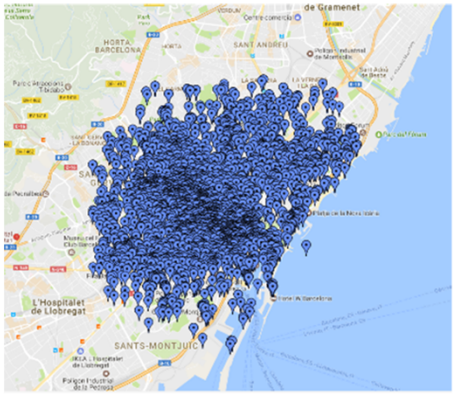


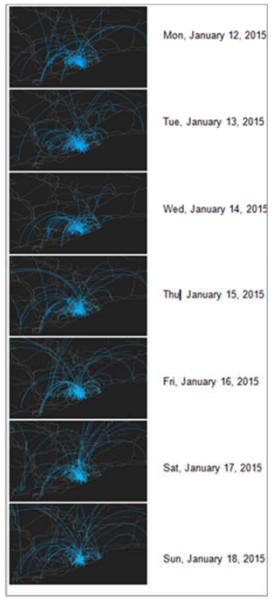
Table 3. Cumulative data for Barcelona, January 13-20, 2017

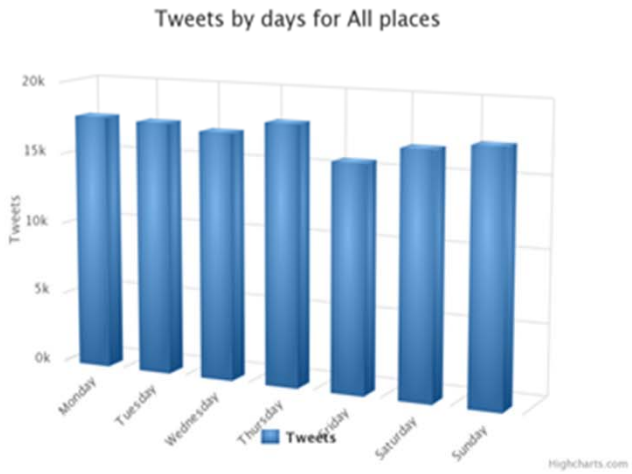
Type of analysis	Barcelona
Number of tweets	117017
Number of users	10412
Number of retweets	20208
Number of likes	51989
Number of applications	45
Number of languages	120

Results

Figure 3. Tweets by day

a. January 12-19, 2015 b. January 13-20, 2017



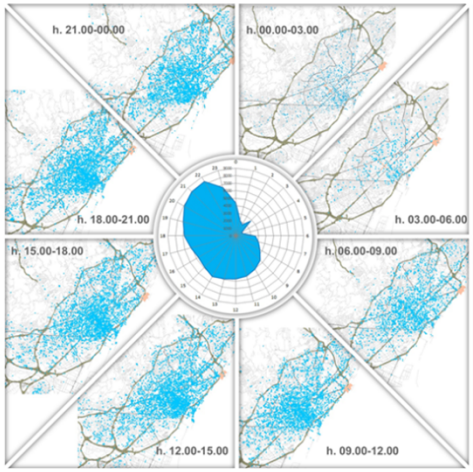


Day	Number of tweets %
Monday	17295 15,00
Tuesday	17122 15,00
Wednesday	16542 14,00
Thursday	16927 14,50
Friday	15681 13,00
Saturday	16533 14,00
Sunday	16917 14,50
Total	117017 100,00
Mean	16717

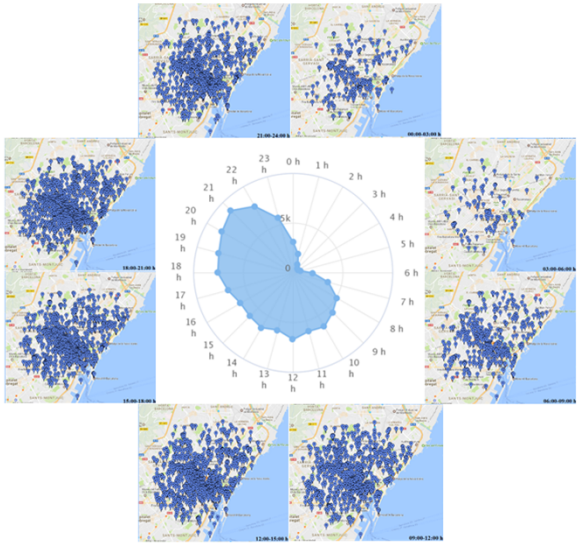
Results

Figure 4. Tweets by time interval, BCN

a. January 12-19, 2015



b. January 13-20, 2017



Interval	Percent of tweets
0-3	4.28%
3-6	1.71%
6-9	7.95%
9-12	13.85%
12-15	17.64%
15-18	18.05%
18-21	20.38%
21-24	16.14%

Results

- EXPERIMENT 2: To analyze the differences between the number of tweets sent depending on the **day of the week**. Pairwise comparisons Chi-Square test. Adjustment for multiple comparisons using Chi-square test with FDR (false discovery rate).

- EXPERIMENT 3: To analyze the differences between the number of tweets sent depending on the **hour of the day.** Pairwise comparisons Chi-Square test. Adjustment for multiple comparisons using Chi-square test with FDR (false discovery rate).

EXPERIMENT 2

Chi-Square Test for Equal Proportions	
Chi-Square	222.8535
DF	6
Pr > ChiSq	<.0001

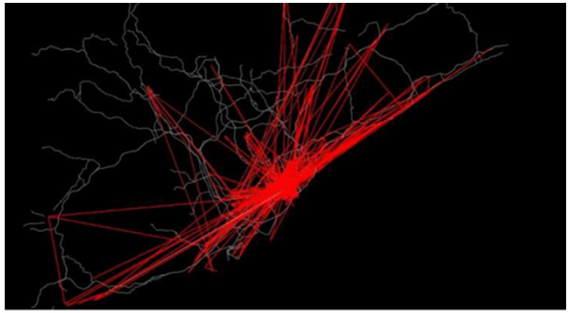
EXPERIMENT 3

Chi-Square Test for Equal Proportions	
Chi-Square	7242.2345
DF	7
Pr > ChiSq	<.0001

Spanish, Catalan and

Barcelona - **Data Mapping**

Galician



Spanish, Catalan and Galician

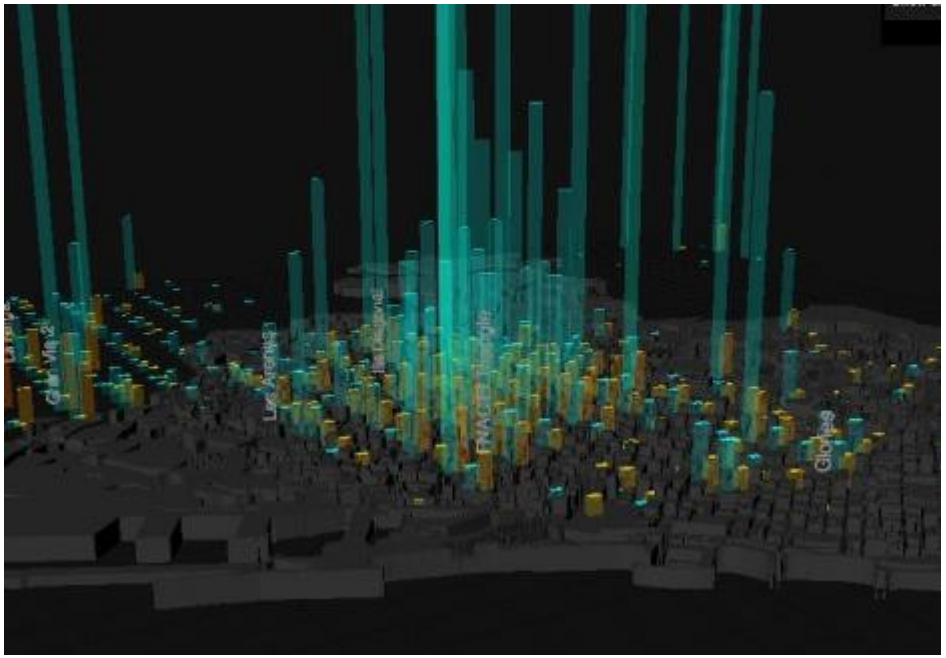
English and European languages

Japanese and Korean languages

Preocupaciones, problemas, preguntas de reflexión para la sala

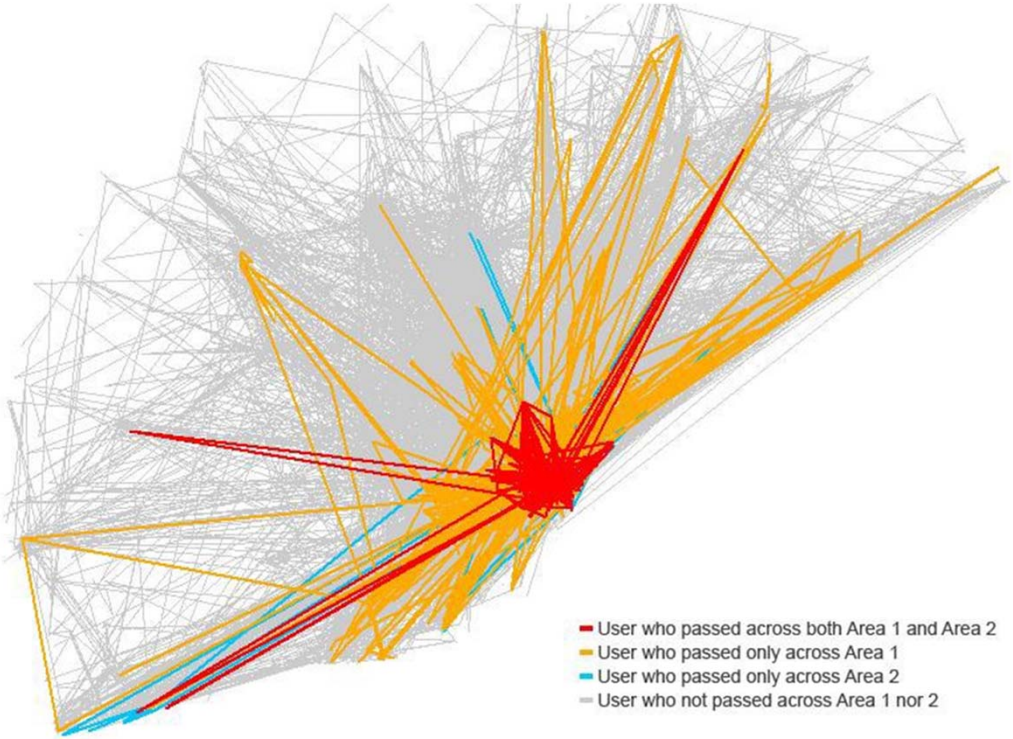
- The main danger of **data-centrism** is that it encourages the idea that whatever the problem, the answer lies in data, not in politics and negotiation.
- The intrusion on **private lives**.
- The **"digital divide," and "Data divide"** (sophisticated analysis is not on the everyone space).
- To make experiment-control using **standard and big data methods** to compare and complement the results.
- Semantic analysis.
- Sesgos en la muestra de datos.

Barcelona: Big Bang data



Gracias!!!

Barcelona - **Data Mapping**



The link between each subsequent tweet[1] for each user over the whole metropolitan area of Barcelona, January 7-19, 2015.

Where the users of Area 1 and Area 2 have been?

1The tracked tweets are only those georeferenced and made public by users

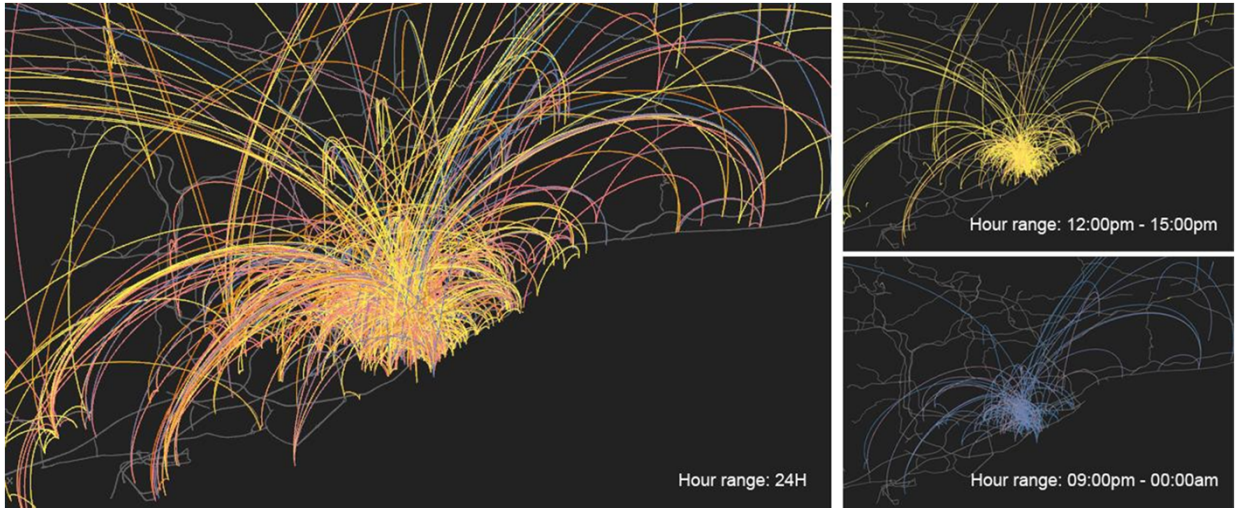
Barcelona - **Data Mapping**

The lines connecting subsequent tweets are superimposed to the road network through the shortest path algorythm. The density of tweets on each arch provides the width of the lines.

Which are the more congested axis?

(railways and tramway tracks are not considered)

January 7-19, 2015.



The link between each subsequent tweet[1] for each user who passed across one or both the two areas to study.
week January 12-19, 2015.

Which are the tweet clusters hour by hour? Identification of the density of activities in the city.