



Exploring Big Data for analysing territorial development

Sofie Jæger

ESPON Scientific Conference

12-13 September 2013

Outline



Introduction

- Big Data
- Twitter

Data Gathering

- Collection
- Analysis
- Limitations & Advantages

Future potentials

- Use in ESPON

Introduction

What is Big Data



Introduction

What is Big Data



- ▶ The term big data, refers to the growing amount of available data

Introduction

What is Big Data



- ▶ The term big data, refers to the growing amount of available data
- ▶ Technological developments enable the gathering and storage of very large amounts of data

Introduction

What is Big Data



- ▶ The term big data, refers to the growing amount of available data
- ▶ Technological developments enable the gathering and storage of very large amounts of data
- ▶ Internet activities are major contributors

Introduction

What is Big Data



- ▶ The term big data, refers to the growing amount of available data
- ▶ Technological developments enable the gathering and storage of very large amounts of data
- ▶ Internet activities are major contributors
 - ▶ e.g. Google searches, Twitter, Facebook, Linkedin, blogs etc



Explored for use in the Flagship project (FP7)



- ▶ Growing social network created in 2006



- ▶ Growing social network created in 2006
- ▶ Tweets are small text messages of max. 140 characters

- ▶ Growing social network created in 2006
- ▶ Tweets are small text messages of max. 140 characters
- ▶ Tweets are published to followers and can be replied to or re-tweetet

Collection

Twitter API



Collection

Twitter API



- Data is collected using the Twitter API

Collection

Twitter API



- ▶ Data is collected using the Twitter API
- ▶ This is done using python scripts. The collection is restricted to tweets that include coordinates and only within a bounding box set for Europe

Collection

Twitter API



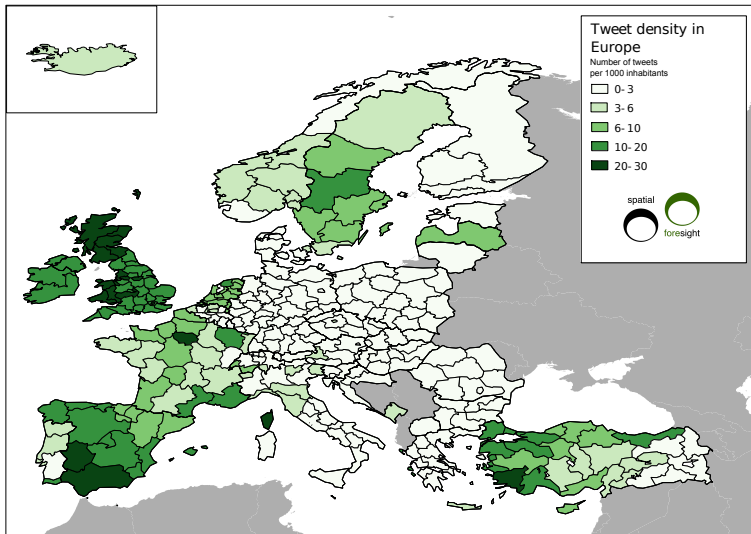
- ▶ Data is collected using the Twitter API
- ▶ This is done using python scripts. The collection is restricted to tweets that include coordinates and only within a bounding box set for Europe
- ▶ Data collected:

- ▶ Data is collected using the Twitter API
- ▶ This is done using python scripts. The collection is restricted to tweets that include coordinates and only within a bounding box set for Europe
- ▶ Data collected:
 - ▶ 'user id', 'place of user', 'tweet id', 'time and date', 'coordinates', 'country code of tweet' and 'language of tweet'. If the tweet is a reply: 'reply id' and 'user id'

- ▶ Data is collected using the Twitter API
- ▶ This is done using python scripts. The collection is restricted to tweets that include coordinates and only within a bounding box set for Europe
- ▶ Data collected:
 - ▶ 'user id', 'place of user', 'tweet id', 'time and date', 'coordinates', 'country code of tweet' and 'language of tweet'. If the tweet is a reply: 'reply id' and 'user id'
- ▶ Data is stored in a SQLite/Spatialite database which can be integrated into QuantumGIS

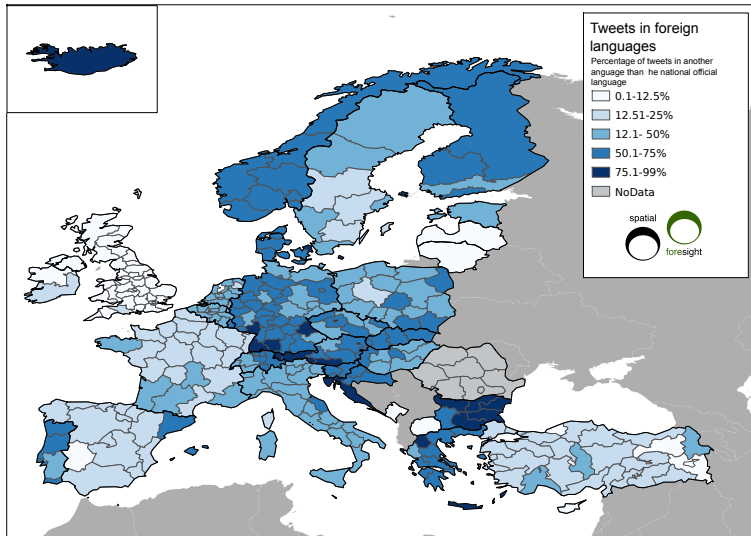
Analysis

Density of tweets



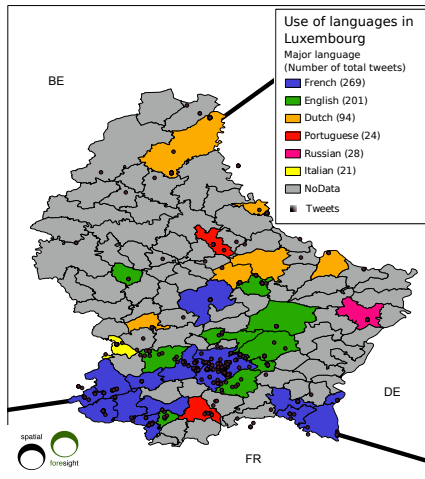
Analysis

Share of foreign languages



Analysis

Share of foreign languages



Potential use

Ideas for other analysis



Potential use

Ideas for other analysis



Flow data

- ▶ Network data based on replies and re-tweets can be used to analyse flows, both within and between countries.

Potential use

Ideas for other analysis



Flow data

- Network data based on replies and re-tweets can be used to analyse flows, both within and between countries.

Keywords and content

- The users index the tweets themselves by using hash tags to define keywords (e.g. #ESPON).

Limitations

What to be aware of when using Twitter data



Limitations

What to be aware of when using Twitter data



- ▶ A good representation of the population in the data is not guaranteed

Limitations

What to be aware of when using Twitter data



- ▶ A good representation of the population in the data is not guaranteed
- ▶ Not all information fields in Twitter are mandatory to fill in

Limitations

What to be aware of when using Twitter data



- ▶ A good representation of the population in the data is not guaranteed
- ▶ Not all information fields in Twitter are mandatory to fill in
- ▶ As the network is based on the needs of the users, additions and changes are occurring continuously

Advantages

In what way can Twitter data be beneficial



Advantages

In what way can Twitter data be beneficial



- ▶ You can get data from parts of the population that are often poorly represented in surveys and questionnaires

Advantages

In what way can Twitter data be beneficial



- ▶ You can get data from parts of the population that are often poorly represented in surveys and questionnaires
- ▶ Data that can be hard to obtain otherwise on a short time scale (e.g. opinions about current events in society) are available from Twitter

Advantages

In what way can Twitter data be beneficial



- ▶ You can get data from parts of the population that are often poorly represented in surveys and questionnaires
- ▶ Data that can be hard to obtain otherwise on a short time scale (e.g. opinions about current events in society) are available from Twitter
- ▶ The dynamics of the network give a very up-to-date picture of the current reality

Use in ESPON

Potentials for future use in ESPON projects



Use in ESPON

Potentials for future use in ESPON projects



- Benefit from more timely data

Use in ESPON

Potentials for future use in ESPON projects



- ▶ Benefit from more timely data
- ▶ Go beyond NUTS3

Use in ESPON

Potentials for future use in ESPON projects



- ▶ Benefit from more timely data
- ▶ Go beyond NUTS3
- ▶ Explore flows between different locations in Europe

Use in ESPON

Potentials for future use in ESPON projects



- ▶ Benefit from more timely data
- ▶ Go beyond NUTS3
- ▶ Explore flows between different locations in Europe
- ▶ Carry out analyses on the content of the information exchanged

#ThanksForYourAttention

#Questions #Discussion



sofie.jaeger@spatialforesight.eu
Twitter@SofieJgr

More details, scripts and data samples:

www.spatialforesight.eu/exploring-big-data.html