Social media sentiment and consumer confidence

Peter Struijs

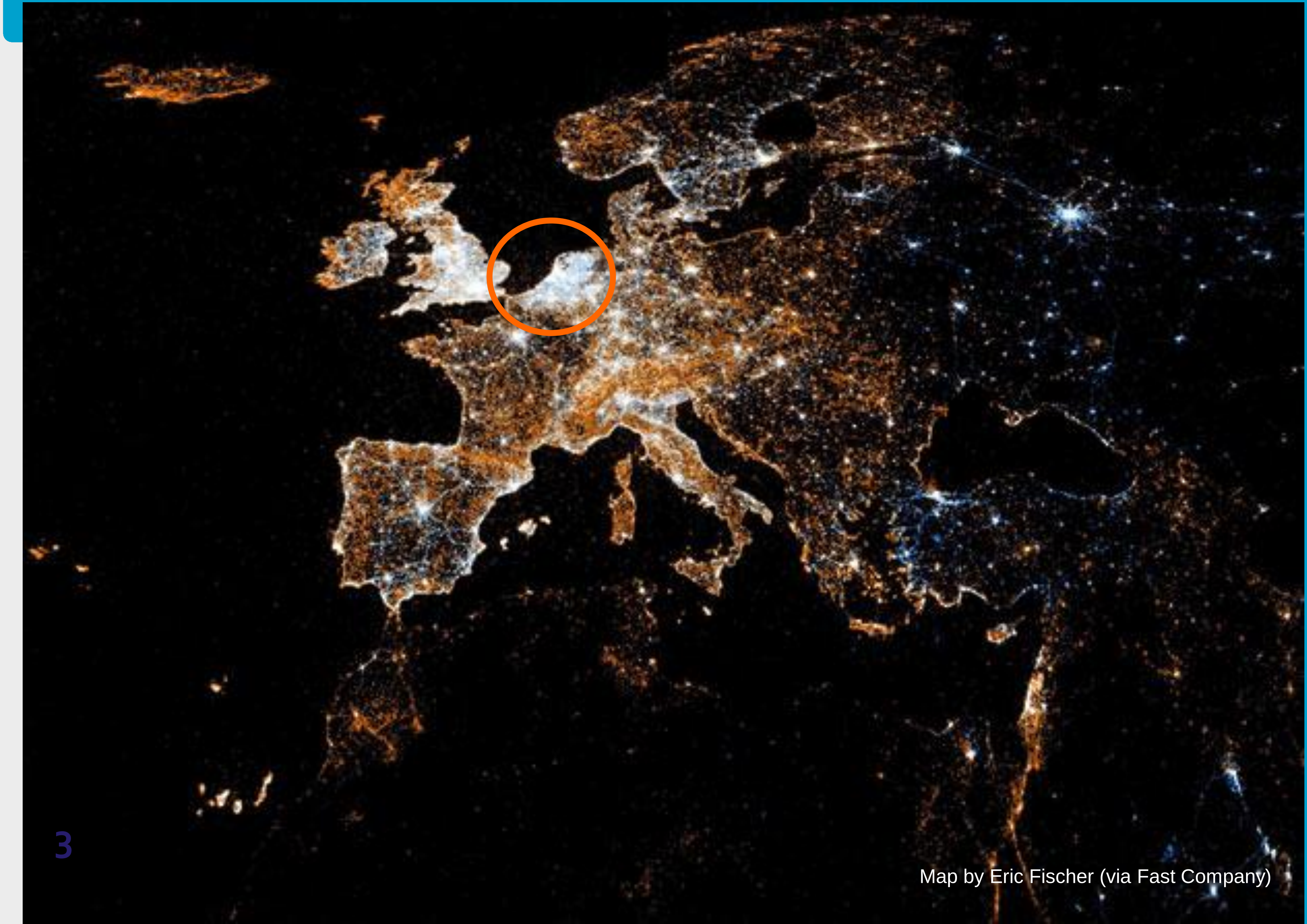
International Conference on Big Data for Official Statistics, Beijing, China, 28 – 30 October 2014

Statistics Netherlands

Research question

Can we replicate the consumer confidence index by only using social media data, while reducing production time?

Social media



Map by Eric Fischer (via Fast Company)

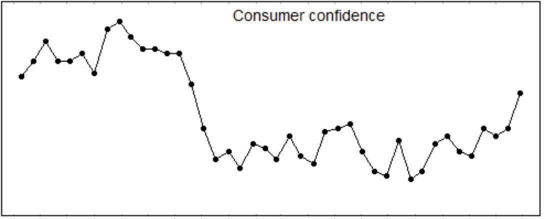
The data

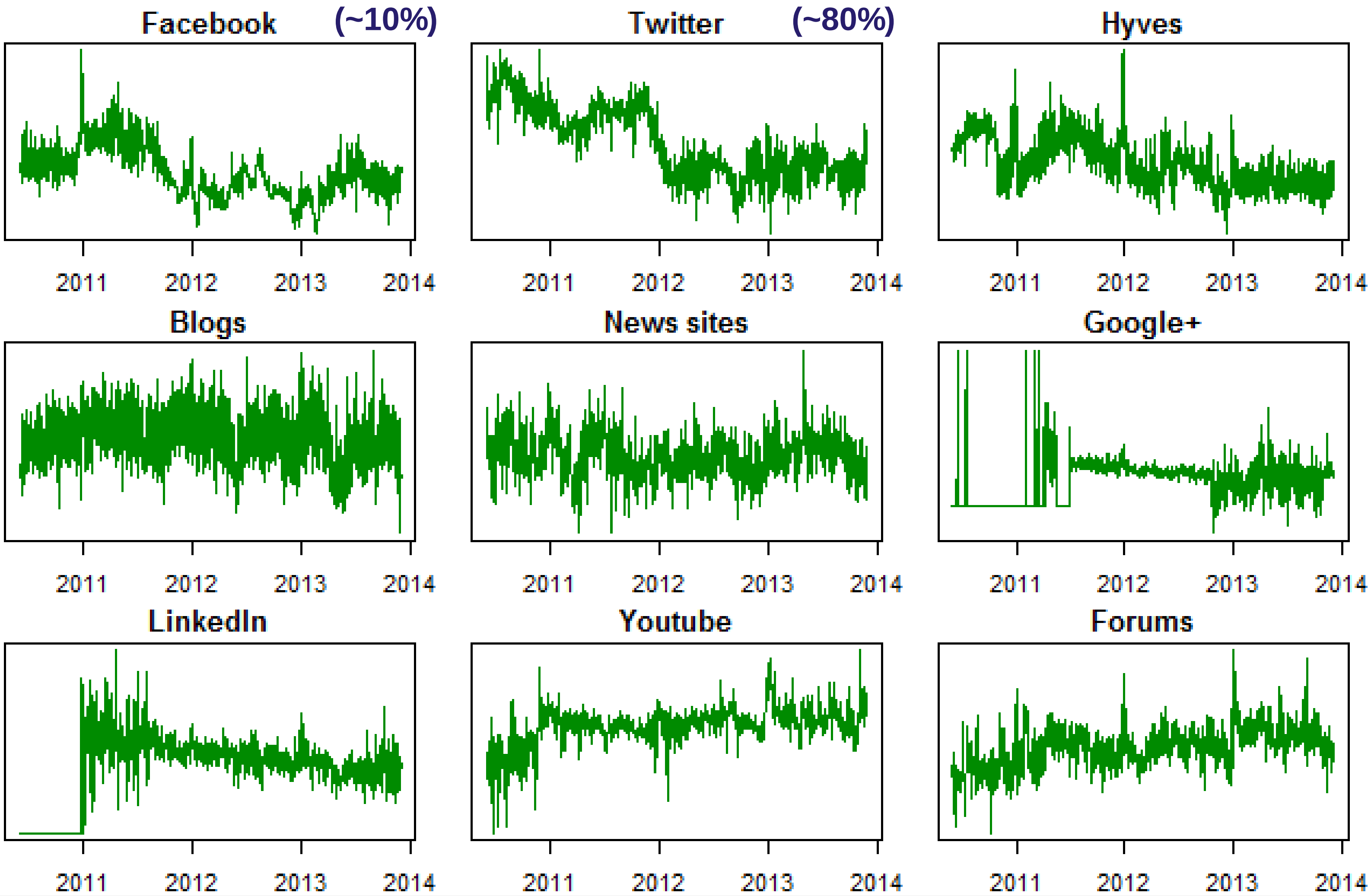
- All social media messages:
  - that are written in Dutch
  - and are public
- These messages are systematically and instantly collected by the Dutch firm Coosto
- Dataset of more than 3.5 billion messages:
  - covering June 2010 till the present
  - between 3-4 million new messages added per day
- Issues:
  - selectivity
  - meaning of the data

Sentiment determination

- ‘Bag of words’ approach:
  - list of Dutch words with their associated sentiment
  - added social media specific words (‘FAIL’, ‘LOL’, ‘OMG’ etc.)
- Use overall score to determine sentiment:
  - is either positive, negative or neutral
- Average sentiment per period (day / week / month)
  - (#positive - #negative)/#total * 100%

Sentiment per platform





Build a model

Idea: Fitting characteristics derived from social media messages to consumer confidence

Success: If correlation can be found that is high and remains high.

Reference:

Daas, P. et al. (2014). Social Media Sentiment and Consumer Confidence. Paper for the Workshop on Using Big Data for Forecasting and Statistics, Frankfurt am Main, Germany.

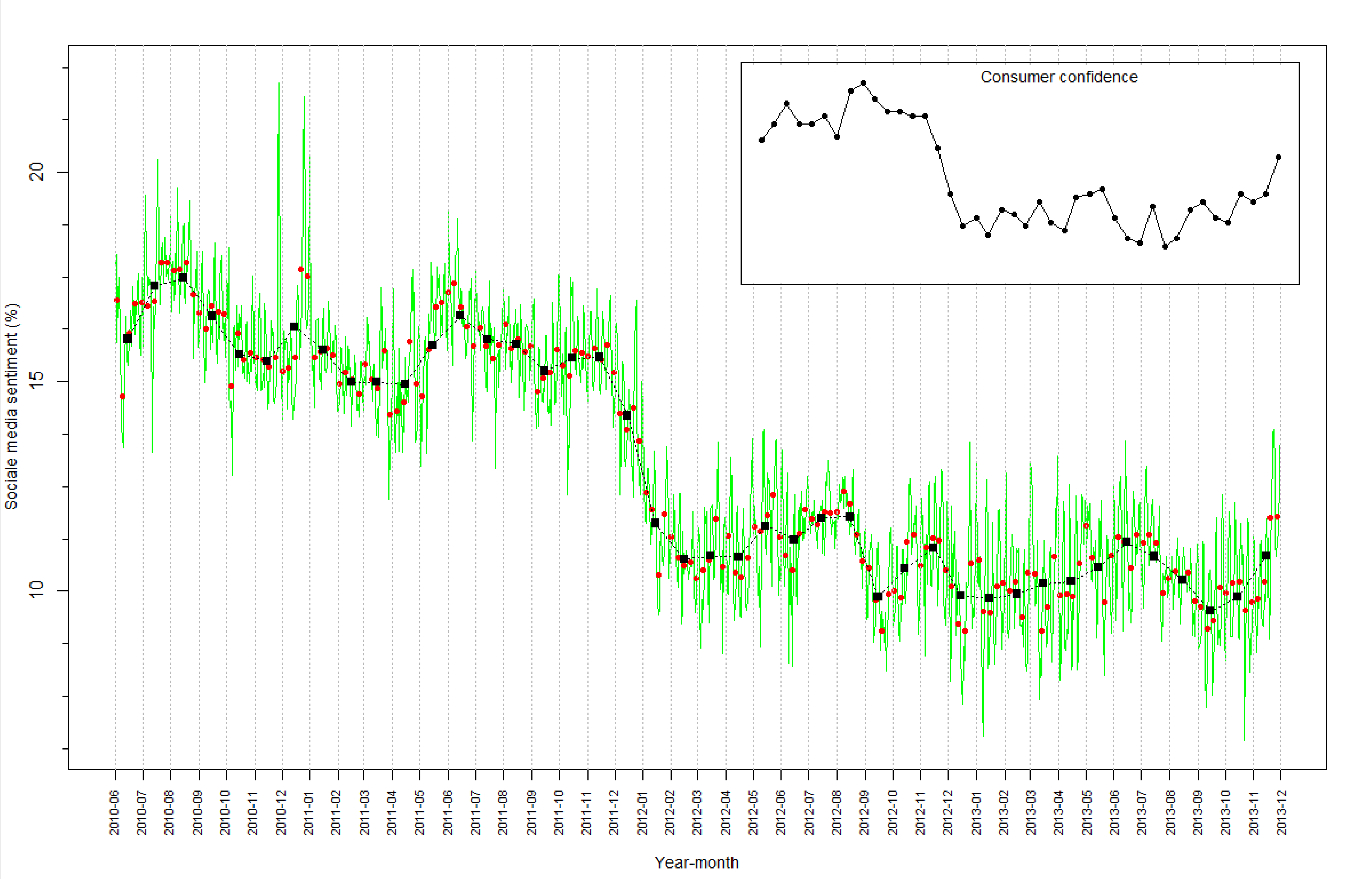


Figure 1. Development of daily, weekly and monthly aggregates of social media sentiment from June 2010 until November 2013, in green, red and black, respectively. In the insert the development of consumer confidence is shown for the identical period.

Results

- High correlation achieved (0.9).
- Changes in consumer confidence preceed changes in sentiment by one week.
- Short processing time, so time-to-market can still be reduced.
- Sentiment index can be produced on a weekly basis.
- To be considered:
  - Use model-based figures as early indicators
  - Reduce sampling of consumer confidence index

Questions

- May official statistics be based on correlations?
- If so, what are the conditions?
- What to do if there is a shock?
- What to do if businesses produce similar information?
- What would be the strategic implications of making statistics in this way?

Lessons learned

- There may be alternatives to population-based estimation methods.
- For research of this type an open mindset is needed.
- A stand-alone research programme may benefit a statistical institute.