On two existing approaches to statistical analysis of social media data

Martina Patone
Li-Chun Zhang

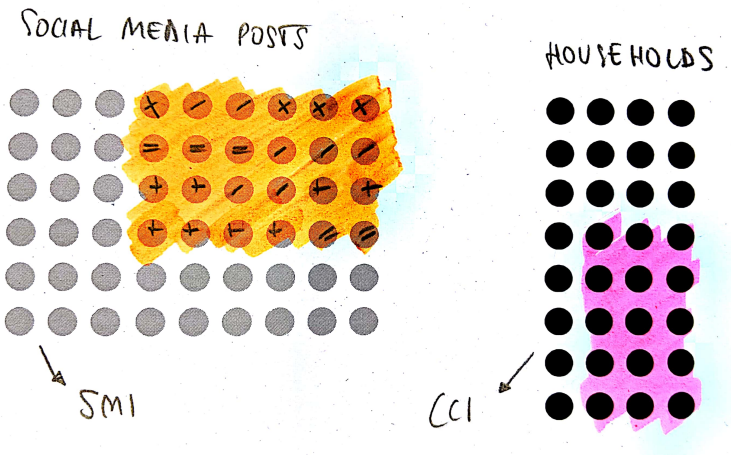
27th October 2018

UNIVERSITY OF Southampton

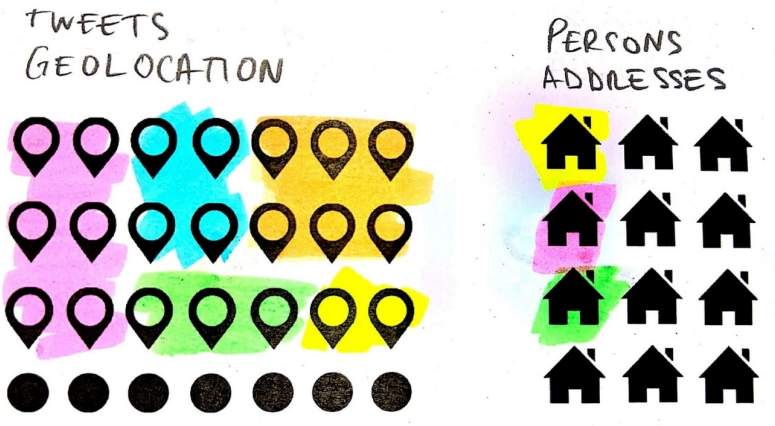
[dstl]

Web Science
Centre for Doctoral Training

The two existing approaches



Approach 1.

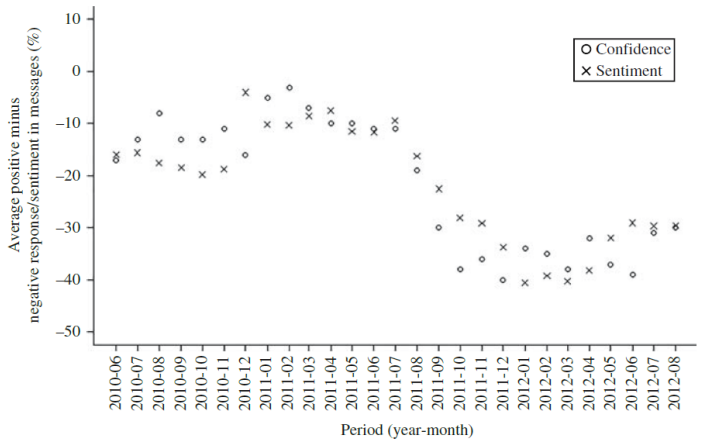


Approach 2.

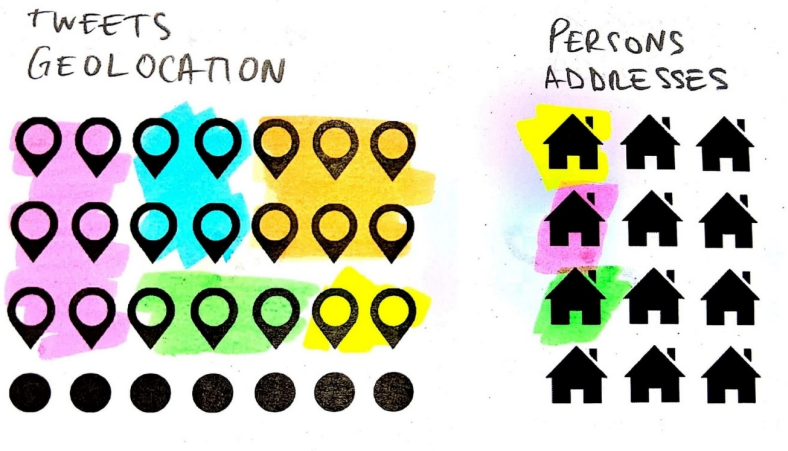
1. The case of the SMI (Daas et al. 2015):
Can the SMI be used to replace the CCI?

2. Inferring users' residency via tweets (Swier et al. 2015):
Can geolocalised tweets be used to extract the residential address of an user?

The two existing approaches



Approach 1.

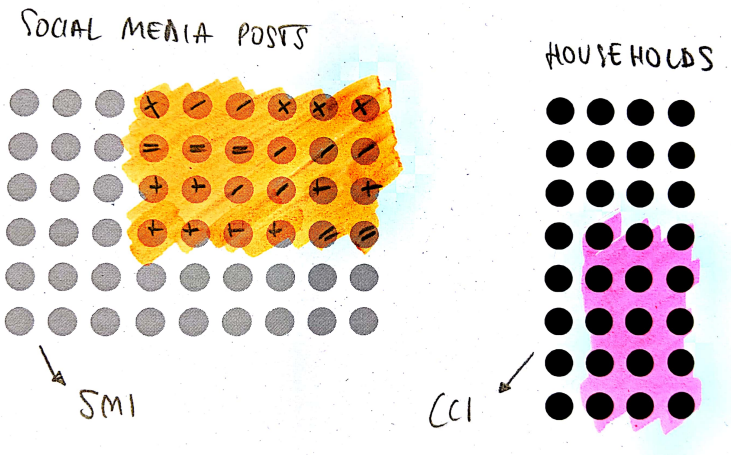


Approach 2.

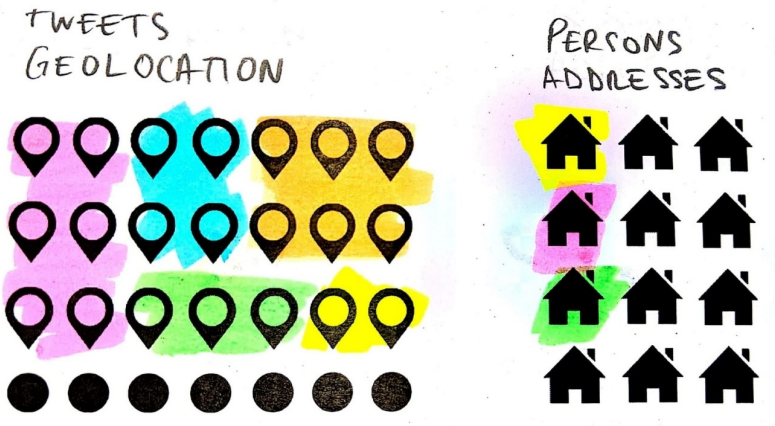
1. The case of the SMI (Daas et al. 2015):
Can the SMI be used to replace the CCI?

2. Inferring users' residency via tweets (Swier et al. 2015):
Can geolocalised tweets be used to extract the residential address of an user?

The two existing approaches



Approach 1.

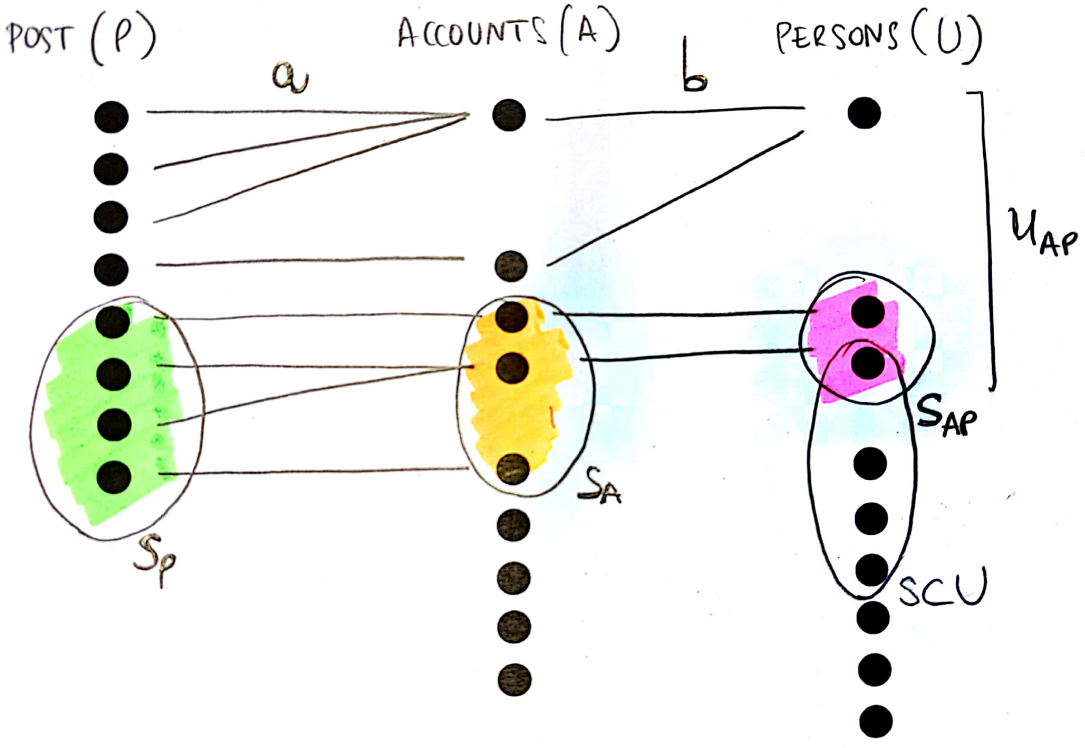


Approach 2.

1. The case of the SMI (Daas et al. 2015):
Can the SMI be used to replace the CCI?

2. Inferring users' residency via tweets (Swier et al. 2015):
Can geolocalised tweets be used to extract the residential address of an user?

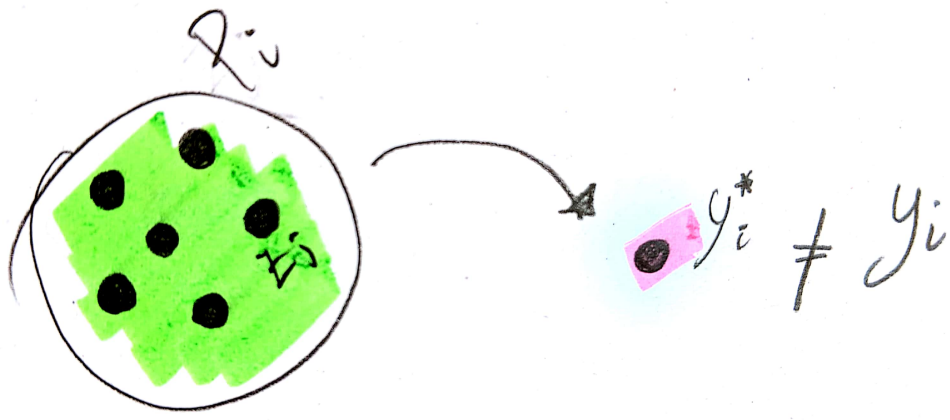
Representation



Connection only between s_P and $s^* = b(a(s_P)) \cap U$

1. **One-phase**: $s_P \subseteq P$ the observed set, U Dutch households;
2. **Two-phase**: $s^* \subseteq U$ the observed set, U UK residents.

Measurement



1. **One-phase**: z_j, $j \in s_P$ the observed sentiment for post j;

2. **Two-phase**: $y_i^* = \tau(z_j, j \in P_i)$ the observed address of the anchor point for person $i \in s^*$.

One-phase: formal interpretation of Daas et al. (2015)

One-phase: use the observed data $(z_j, s_{P,t})$ to aim at the same parameter $\theta = \theta(y_U)$

$$\theta = \xi(\mathbf{z_{s_{P,t}}}) = \theta(\mathbf{y_U})$$

$$\mathsf{SMI}_t = \xi(z_{s_{P,t}})$$
$$\mathsf{SMI}_t = \xi_t + d_t$$
$$E(d_t) = 0, V(d_t) = \eta_t^2 (\approx 0)$$

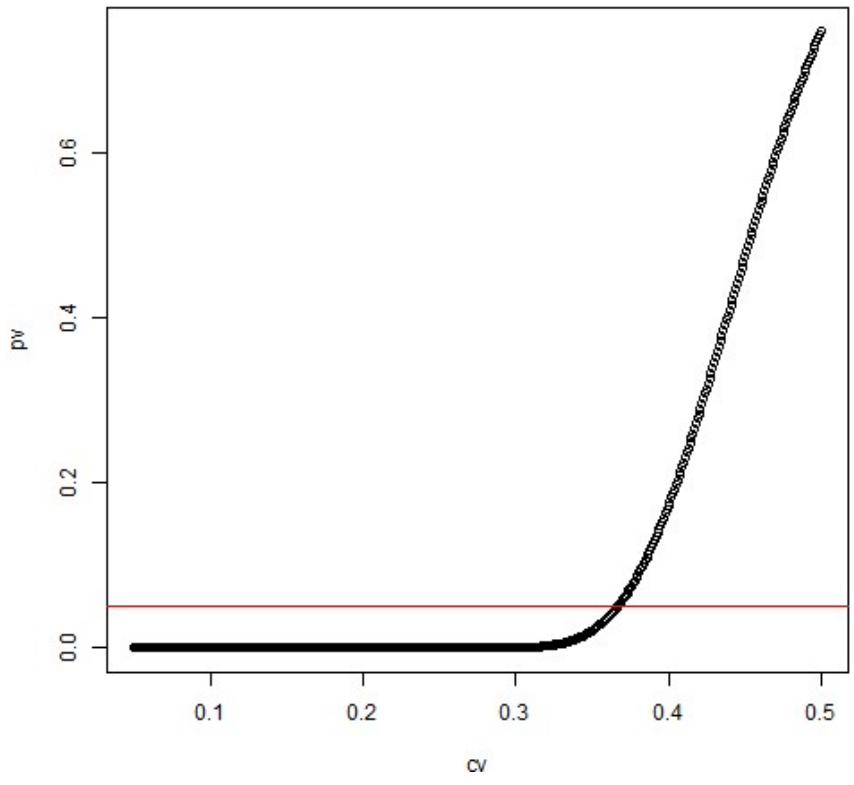
$$\mathsf{CCI}_t = \theta(y_{s_t})$$
$$\mathsf{CCI}_t = \theta_t + e_t$$
$$e_t \sim \mathcal{N}(0, \sigma_t^2)$$

Can the SMI replace the CCI?: $\theta_t = \xi_t$

One-phase: statistical validation

Test: $H_0 : \theta_t - \xi_t = \mu$ *vs.* $H_1 : \theta_t - \xi_t \neq \mu$;

Under H_0: $X_t = \mathsf{CCI}_t - \mathsf{SMI}_t = \mu + e_t$ with $e_t \sim \mathcal{N}(0, \sigma_t^2)$;



p-value exceeds 0.05 for $cv > 0.367$

Two-phase: formal interpretation of Swier et al. (2015)

Two-phase: transform the social media dataset (z_j, s_P) in a pseudo-survey dataset (y_i^*, s^*)

$$s_P \xrightarrow{a} s_A \xrightarrow{b} s^* \subset U \qquad \textit{representation}$$

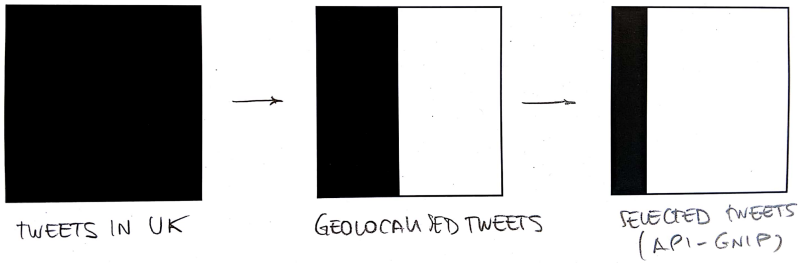
$$z_j \rightarrow y_i^* (\neq y_i),\ j \in P_i \qquad \textit{measurement}$$

Statistical analysis is performed on (y_i^*, s^*)

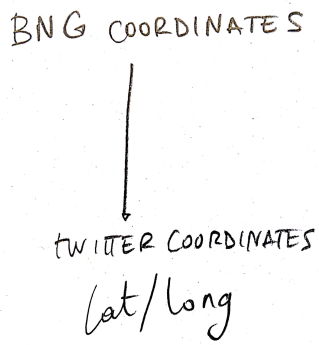
Two-phase: data quality

1^{st} **phase**: social media dataset

Representation

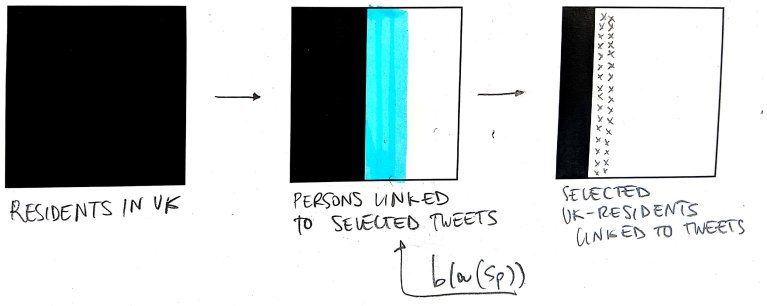


Measurement

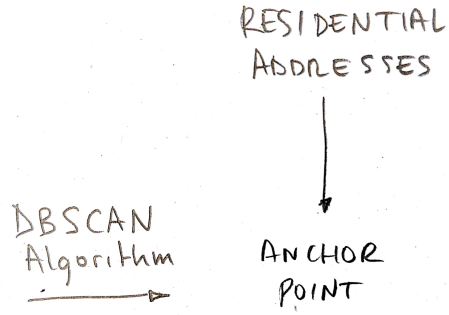


2^{nd} **phase**: pseudo-survey dataset

Representation



Measurement



One-phase and Two-phase: discussion

	One-phase (z_j, s_P)	Two-phase (y_i^*, s^*)
(z_j, P)	non-probability sample $s_p \subset P$ unknown selection	not of interest
(y_i^*, s^*)	need weighting	feasible
(y_i, s^*)	need weighting measurement consideration	feasible measurement consideration
(y_i^*, U)	test parameters need survey data	non-probability sample $s^* \subset U$ unknown selection
(y_i, U)	test parameters need survey data	non-probability sample $s^* \subset U$ unknown selection measurement consideration