Population bias in geotagged tweets

Workshop on Standards and Practices in Large-Scale Social Media Research
9th International AAAI Conference on Weblogs and Social Media (ICWSM '15)
May 26-29, 2015, Oxford, UK

Momin M. Malik
Hemank Lamba
Constantine Nakos
Jürgen Pfeffer

Institute for Software Research
School of Computer Science
Carnegie Mellon University

March 26, 2015

Slides at `http://mominmalik.com/Malik_ICWSM2015_slides.pdf`

Motivation

1. All maps of geotagged tweets look like maps of population density
2. LOTS of studies use geotagged tweets
3. Need for large-scale, multivariate, statistical work

Data

1. Geotags in tweets
2. US Census
3. Mobile users

Model

1. A statistical test for random distribution over population
2. Validating the null model
3. Spatial autocorrelation and spatial errors
4. Model specification

Results

1. Test of null hypothesis
2. Spatial autocorrelation
3. Spatial errors model results

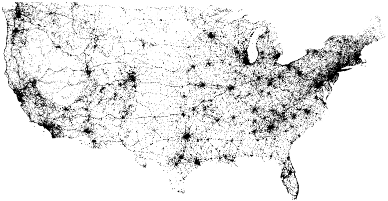
Discussion

1. Limitations
2. Take-away
3. Future work

All maps of geotagged tweets look like maps of population density

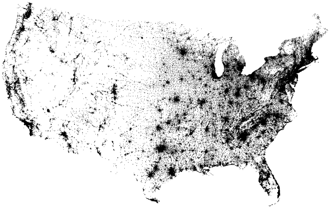
All maps of geotagged tweets look like maps of population density

Tweets:



Adapted from 'Contiguous United States geotag map (2009)' by Eric Fischer (https://www.flickr.com/photos/walkingsf/5985800498)

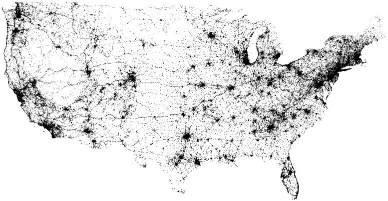
Population:



Population density in 2010 US Census. Adapted from 'Nighttime Population Distribution Wall Map' by Geography Division, U.S. Department of Commerce / Economics and Statistics Administration / U.S. Census Bureau. Each square represents 1,000 people.

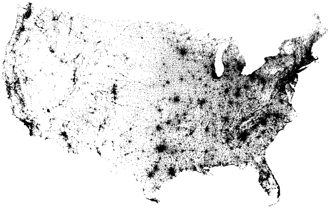
All maps of geotagged tweets look like maps of population density

Tweets:



Adapted from ‘Contiguous United States geotag map (2009)’ by Eric Fischer (https://www.flickr.com/photos/walkingsf/5985800498)

Population:



Population density in 2010 US Census. Adapted from ‘Nighttime Population Distribution Wall Map’ by Geography Division, U.S. Department of Commerce / Economics and Statistics Administration / U.S. Census Bureau. Each square represents 1,000 people.

We’d like to correct for population density.

LOTS of studies use geotagged tweets

...and for a lot of things!

LOTS of studies use geotagged tweets

...and for a lot of things!

- Mobility (Yuan et al., 2013; Cho et al., 2011);
- Urban life (Doran et al., 2013; Frias-Martinez et al., 2012);
- Transportation (Wang et al., 2014);
- Natural disasters, crises, and disaster response (Morstatter et al., 2014; Lin and Margolin, 2014; Shelton et al., 2014; Sylvester et al., 2014; Kumar et al., 2014);
- Public health (Sylvester et al., 2014; Nagar et al., 2014; Ghosh and Guha, 2013)
- Language (Hong et al., 2012; Eisenstein et al., 2010; Kinsella et al., 2011);
- Discourse (Leetaru et al., 2013);
- Information diffusion/flows (Kamath et al., 2013; van Liere, 2010);
- Emotion (Mitchell et al., 2013);
- Social ties (Stephens and Poorthuis, 2014; Takhteyev et al., 2012; Cho et al., 2011)

LOTS of studies use geotagged tweets

...and for a lot of things!

- Mobility (Yuan et al., 2013; Cho et al., 2011);
- Urban life (Doran et al., 2013; Frias-Martinez et al., 2012);
- Transportation (Wang et al., 2014);
- Natural disasters, crises, and disaster response (Morstatter et al., 2014; Lin and Margolin, 2014; Shelton et al., 2014; Sylvester et al., 2014; Kumar et al., 2014);
- Public health (Sylvester et al., 2014; Nagar et al., 2014; Ghosh and Guha, 2013)
- Language (Hong et al., 2012; Eisenstein et al., 2010; Kinsella et al., 2011);
- Discourse (Leetaru et al., 2013);
- Information diffusion/flows (Kamath et al., 2013; van Liere, 2010);
- Emotion (Mitchell et al., 2013);
- Social ties (Stephens and Poorthuis, 2014; Takhteyev et al., 2012; Cho et al., 2011)

Implicit assumption that geotagged tweets tell us about the larger world. But do they?

Need for large-scale, multivariate, statistical work

Need for large-scale, multivariate, statistical work

- Mislove et al. (2011):
  - Method: connected user-specified ‘location’ field to county-level US Census data (data prior to geotags in Twitter)
  - Findings: overrepresentation of populous counties and cities with large white populations; underrepresentation of midwest, southwest Hispanic populations, south and midwestern black populations

Need for large-scale, multivariate, statistical work

- Mislove et al. (2011):
  - Method: connected user-specified ‘location’ field to county-level US Census data (data prior to geotags in Twitter)
  - Findings: overrepresentation of populous counties and cities with large white populations; underrepresentation of midwest, southwest Hispanic populations, south and midwestern black populations
- Hecht and Stephens (2014):
  - Method: connects 56.7m tweets across US from 1.6m users over 25 days in 2013 to data from US Census and other federal agencies
  - Findings: Urban areas have 2.7 to 3.5 times more geotagged tweet users than we would expect

Need for large-scale, multivariate, statistical work

- Mislove et al. (2011):
  - Method: connected user-specified 'location' field to county-level US Census data (data prior to geotags in Twitter)
  - Findings: overrepresentation of populous counties and cities with large white populations; underrepresentation of midwest, southwest Hispanic populations, south and midwestern black populations
- Hecht and Stephens (2014):
  - Method: connects 56.7m tweets across US from 1.6m users over 25 days in 2013 to data from US Census and other federal agencies
  - Findings: Urban areas have 2.7 to 3.5 times more geotagged tweet users than we would expect
- Longley et al. (2015):
  - Method: gets profiles of geotag tweet users in greater London area, uses forename-surname matching to identify gender, age and ethnicity to compare to (2011 UK) Census data
  - Findings: overrepresentation of young males and white British users; underrepresentation of middle-aged/older females, and of South Asian, West Indian, and Chinese users

Need for large-scale, multivariate, statistical work

- Mislove et al. (2011):
  - Method: connected user-specified ‘location’ field to county-level US Census data (data prior to geotags in Twitter)
  - Findings: overrepresentation of populous counties and cities with large white populations; underrepresentation of midwest, southwest Hispanic populations, south and midwestern black populations
- Hecht and Stephens (2014):
  - Method: connects 56.7m tweets across US from 1.6m users over 25 days in 2013 to data from US Census and other federal agencies
  - Findings: Urban areas have 2.7 to 3.5 times more geotagged tweet users than we would expect
- Longley et al. (2015):
  - Method: gets profiles of geotag tweet users in greater London area, uses forename-surname matching to identify gender, age and ethnicity to compare to (2011 UK) Census data
  - Findings: overrepresentation of young males and white British users; underrepresentation of middle-aged/older females, and of South Asian, West Indian, and Chinese users
- Pew survey of location services (Zickuhr, 2013):
  - n=1,178; ‘geosocial’ n=141; Twitter ‘geosocial’ n=1
  - Low*est* and middle income use most; low*er* use less, high use least; more 18-26 year-olds, more Hispanic users

Motivation

1. All maps of geotagged tweets look like maps of population density
2. LOTS of studies use geotagged tweets
3. Need for large-scale, multivariate, statistical work

Data

1. Geotags in tweets
2. US Census
3. Mobile users

Model

1. A statistical test for random distribution over population
2. Validating the null model
3. Spatial autocorrelation and spatial errors
4. Model specification

Results

1. Test of null hypothesis
2. Spatial autocorrelation
3. Spatial errors model results

Discussion

1. Limitations
2. Take-away
3. Future work

Geotags in tweets

What geotags look like:


```
https://api.twitter.com/1.1/statuses/show/
123456789012345678.json

{
    "created_at": "Wed Apr 01 00:47:05
                +00002015",
    "text": "This view tho \uE106\uE00E,
    "user": {
        "followers_count": 36000,
        "friends_count": 25000,
        "geo_enabled": true,
    },
    "geo": {
        "type": "Point",
        "coordinates":
        [36.11570625,-115.17407114]
    }
}
```

Geotags in tweets

About geotags:

- Latitude and longitude to the ten thousandth of a degree
- Automatically generated once user enables (unlike 'location')
- Accessible via Twitter's Streaming API
  - We used geobox [124.7625, 66.9326]W $\times$ [24.5210, 49.3845]N
  - From April 1 to July 1, 2013, collected **144,877,685** geotagged tweets, representing **2,612,876** unique twitter handles.

Geotags in tweets

About geotags:

- Latitude and longitude to the ten thousandth of a degree
- Automatically generated once user enables (unlike 'location')
- Accessible via Twitter's Streaming API
  - We used geobox [124.7625, 66.9326]W $\times$ [24.5210, 49.3845]N
  - From April 1 to July 1, 2013, collected **144,877,685** geotagged tweets, representing **2,612,876** unique twitter handles.

Twitter users sample the population. But the (public) API samples Twitter.

- API queries for which matches exceed 1% of all tweets are non-randomly sampled (Morstatter et al., 2013)
- 1.23% of total volume of tweets are geotagged (Liu et al., 2014)
- US accounts for $\sim$22% of all geotagged tweets (Morstatter et al., 2013)
- 22% $\times$ 1.23% $<$ 1%, so we believe we have everything

US Census

Map data:

- Data are available per “block group”
  - Block $\subset$ **Block Group** $\subset$ Tract $\subset$ County $\subset$ State
  - 220,334 block groups (215,798 in contiguous US).
  - Unique 12-digit “FIPS” codes
  - 0.002 square miles to over 7,500 square miles. Designed to have comparable populations (300 to 6,000 people).
- Block group “shape files” (map file format) are available from Census
- Use Python `shapely` package to place tweets in block groups.

US Census

Map data:

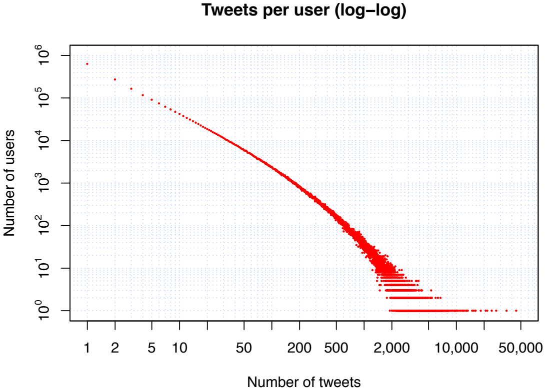
- Data are available per "block group"
  - Block $\subset$ **Block Group** $\subset$ Tract $\subset$ County $\subset$ State
  - 220,334 block groups (215,798 in contiguous US).
  - Unique 12-digit "FIPS" codes
  - 0.002 square miles to over 7,500 square miles. Designed to have comparable populations (300 to 6,000 people).
- Block group "shape files" (map file format) are available from Census
- Use Python `shapely` package to place tweets in block groups.

Demographic data:

- 2010 Decennial Census
  - Tries to count everybody, not sample
  - Population, Race, Gender, Age, Urban, etc. by block group
- American Community Survey (ACS)
  - Done at intervals of 1 year (2013), 3 years (2011-2013), 5 years (2009-2013)
  - From sampling and inference; only 5-year ACS has block groups
  - Use this to get median income, which is not in the Census

Assigning unique locations to mobile users

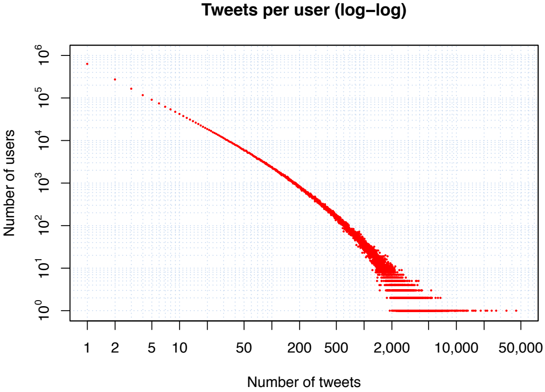
The distribution of tweets per user is, as usual, highly skewed:



Want to use the number of *users*, rather than the number of tweets. Use 'plurality' rule to assign users uniquely (Hecht and Stephens, 2014).

Assigning unique locations to mobile users

The distribution of tweets per user is, as usual, highly skewed:



Want to use the number of *users*, rather than the number of tweets. Use ‘plurality’ rule to assign users uniquely (Hecht and Stephens, 2014).

We also try filtering out users with below 5 geotagged tweets, and users below 10 geotagged tweets. This yields subtly different results.

Motivation

1. All maps of geotagged tweets look like maps of population density
2. LOTS of studies use geotagged tweets
3. Need for large-scale, multivariate, statistical work

Data

1. Geotags in tweets
2. US Census
3. Mobile users

Model

1. A statistical test for random distribution over population
2. Validating the null model
3. Spatial autocorrelation and spatial errors
4. Model specification

Results

1. Test of null hypothesis
2. Spatial autocorrelation
3. Spatial errors model results

Discussion

1. Limitations
2. Take-away
3. Future work

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about:

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about: the quantity U is proportional to population P,

$$U = \alpha P$$

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about: the quantity U is proportional to population P, plus some mean-zero noise term ε

$$U = \alpha P + \varepsilon$$

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about: the quantity U is proportional to population P, plus some mean-zero noise term ε that is proportional to the population.

$$U = \alpha P + \varepsilon P$$

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about: the quantity U is proportional to population P, plus some mean-zero noise term ε that is proportional to the population.

$$U = \alpha P + \varepsilon P$$

Take a log transformation to stabilize the variance.

$$\log U = \log\left(\alpha P + \varepsilon P\right)$$

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about: the quantity U is proportional to population P, plus some mean-zero noise term ε that is proportional to the population.

$$U = \alpha P + \varepsilon P$$

Take a log transformation to stabilize the variance.

$$\log U = \log\left(\alpha P + \varepsilon P\right) = \log\left(\alpha P\left(1 + \frac{\varepsilon}{\alpha}\right)\right)$$

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about: the quantity U is proportional to population P, plus some mean-zero noise term ε that is proportional to the population.

$$U = \alpha P + \varepsilon P$$

Take a log transformation to stabilize the variance.

$$\begin{aligned}\log U = \log(\alpha P + \varepsilon P) &= \log\left(\alpha P\left(1 + \frac{\varepsilon}{\alpha}\right)\right) \\ &= \log\alpha + \log P + \log\left(1 + \frac{\varepsilon}{\alpha}\right)\end{aligned}$$

A statistical test for random distribution over population

We want to test if users are randomly distributed over the US population.

What would it look like for something to be randomly distributed over the population?

How about: the quantity U is proportional to population P, plus some mean-zero noise term ε that is proportional to the population.

$$U = \alpha P + \varepsilon P$$

Take a log transformation to stabilize the variance.

$$\begin{aligned}\log U = \log(\alpha P + \varepsilon P) &= \log\left(\alpha P\left(1 + \frac{\varepsilon}{\alpha}\right)\right) \\ &= \log \alpha + \log P + \log\left(1 + \frac{\varepsilon}{\alpha}\right)\end{aligned}$$

Note that $\log\left(1 + \frac{\varepsilon}{\alpha}\right)$, very conveniently, now has mean zero. Call this ε'.

A statistical test for random distribution over population

$$\log U = \log \alpha + \log P + \varepsilon' \tag{1}$$

A statistical test for random distribution over population

$$\log U = \log \alpha + \log P + \varepsilon' \tag{1}$$

Now, what if we were to fit a linear model of the log population against the log number of users?

A statistical test for random distribution over population

$$\log U = \log \alpha + \log P + \varepsilon' \tag{1}$$

Now, what if we were to fit a linear model of the log population against the log number of users?

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{2}$$

A statistical test for random distribution over population

$$\log U = \log \alpha + \log P + \varepsilon' \tag{1}$$

Now, what if we were to fit a linear model of the log population against the log number of users?

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{2}$$

If eqn. (1) described the true data-generating process,

A statistical test for random distribution over population

$$\log U = \log \alpha + \log P + \varepsilon' \tag{1}$$

Now, what if we were to fit a linear model of the log population against the log number of users?

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{2}$$

If eqn. (1) described the true data-generating process, $e^{\hat{\beta}_0}$ would be an estimator of α,

A statistical test for random distribution over population

$$\log U = \log \alpha + 1 \log P + \varepsilon' \tag{1}$$

Now, what if we were to fit a linear model of the log population against the log number of users?

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{2}$$

If eqn. (1) described the true data-generating process, $e^{\hat{\beta}_0}$ would be an estimator of α, and $\hat{\beta}_1$ should be 1.

A statistical test for random distribution over population

$$\log U = \log \alpha + 1 \log P + \varepsilon' \tag{1}$$

Now, what if we were to fit a linear model of the log population against the log number of users?

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{2}$$

If eqn. (1) described the true data-generating process, $e^{\hat{\beta}_0}$ would be an estimator of α, and $\hat{\beta}_1$ should be 1.

This gives us a concrete null hypothesis that we can test, $H_0 : \beta_1 = 1$. We will fit the model of eqn (2) and see if we can reject this null.

Validating the null model

Is this a good null model?

Validating the null model

Is this a good null model?

Validate by trying this model on the number of males in the population. How good is the fit? We can calculate the true male ratio in our data, .4915; will the model capture this? And will our model give a coefficient 1?

Validating the null model

Is this a good null model?

Validate by trying this model on the number of males in the population. How good is the fit? We can calculate the true male ratio in our data, .4915; will the model capture this? And will our model give a coefficient 1?

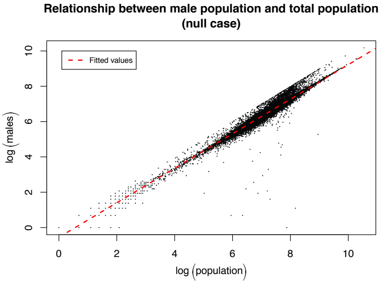
	Dependent variable:	
	log(males)	s.e.
Intercept	-0.731*	(0.002)
log(pop)	1.002*	(0.0003)
Obs.	215,476	
R^2	0.975	
Resid. S.E.	0.081	
F Stat	8,350,415*	
Note:		$^*p<2.2\times10^{-16}$

Validating the null model

Is this a good null model?

Validate by trying this model on the number of males in the population. How good is the fit? We can calculate the true male ratio in our data, .4915; will the model capture this? And will our model give a coefficient 1?

	Dependent variable:	
	log(males)	s.e.
Intercept	-0.731*	(0.002)
log(pop)	1.002*	(0.0003)
Obs.	215,476	
R^2	0.975	
Resid. S.E.	0.081	
F Stat	8,350,415*	
Note:	$^*p<2.2\times10^{-16}$	

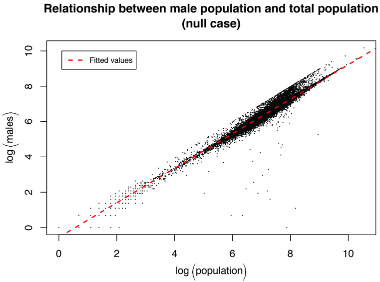


Validating the null model

Is this a good null model?

Validate by trying this model on the number of males in the population. How good is the fit? We can calculate the true male ratio in our data, .4915; will the model capture this? And will our model give a coefficient 1?

	Dependent variable:	
	log(males)	s.e.
Intercept	-0.731*	(0.002)
log(pop)	1.002*	(0.0003)
Obs.	215,476	
R^2	0.975	
Resid. S.E.	0.081	
F Stat	8,350,415*	
Note:	*p<2.2×10^−16	



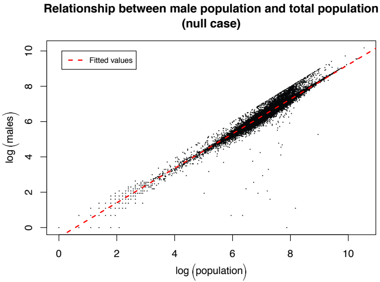
A 95% CI for α is $[e^{-0.729}, e^{-0.733}] = [.4914, .4962]$.

Validating the null model

Is this a good null model?

Validate by trying this model on the number of males in the population. How good is the fit? We can calculate the true male ratio in our data, .4915; will the model capture this? And will our model give a coefficient 1?

	Dependent variable:	
	log(males)	s.e.
Intercept	-0.731*	(0.002)
log(pop)	1.002*	(0.0003)
Obs.	215,476	
R^2	0.975	
Resid. S.E.	0.081	
F Stat	8,350,415*	
Note:	*p$<2.2\times10^{-16}$	



A 95% CI for α is $[e^{-0.729}, e^{-0.733}] = [.4914, .4962]$.

A 95% CI for the coefficient of log population is $[1.001, 1.003]$; 1 is outside this, but this is without accounting for spatial autocorrelation.

Spatial autocorrelation

Adjacent geographic units are not independent; there is *spatial autocorrelation*. Measure this with Moran's I,

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_i (X_i - \overline{X})^2} \tag{3}$$

which is empirical covariance between adjacent units, appropriately normalized. $[w_{ij}] = \mathbf{W}$ is an $n \times n$ matrix of weights (adjacencies between geographic units).

Spatial autocorrelation

Adjacent geographic units are not independent; there is *spatial autocorrelation*. Measure this with Moran's I,

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_i (X_i - \overline{X})^2} \tag{3}$$

which is empirical covariance between adjacent units, appropriately normalized. $[w_{ij}] = \mathbf{W}$ is an $n \times n$ matrix of weights (adjacencies between geographic units).

W is a substantive modeling choice (Gaetan and Guyon, 2012), but if no prior knowledge, test out different options (Anselin et al., 2007). We try:

- Rook contiguity; shared edge only (can normalize rows by row sum)
- Queen contiguity; shared edge or vertex (again, can normalize rows)
- k-nearest neighbors (using block group centroid) for $k = 2, ..., 8$.

Look for spatial autocorrelation in the residuals of a linear model instead of in individual variables (Anselin and Rey, 1991).

Spatial errors model

With **W**, correct for spatial autocorrelation in a *spatial errors model*, a type of simultaneous autoregressive (SAR) model.

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{4}$$

$$\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \tag{5}$$

where **u** is the vector of correlated residuals, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ are the uncorrelated error terms, and λ is the strength of the spatial autocorrelation (Anselin, 2002). Equivalently,

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{I} - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon} \tag{6}$$

Spatial errors model

With **W**, correct for spatial autocorrelation in a *spatial errors model*, a type of simultaneous autoregressive (SAR) model.

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{4}$$

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \tag{5}$$

where **u** is the vector of correlated residuals, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ are the uncorrelated error terms, and λ is the strength of the spatial autocorrelation (Anselin, 2002). Equivalently,

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{I} - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon} \tag{6}$$

Implemented in R package `spdep` (Bivand and Piras, 2015; Bivand et al., 2013a,b). Fits by finding the log determinant of $|\mathbf{I} - \lambda\mathbf{W}|$; infeasible for $n = 215,798$.

Is Cholesky decomposition option, but that requires symmetric matrix, limiting our choices of **W**.

Model specification

- Following previous literature, we include covariates for:
  - Black population, Asian population, Latino/Hispanic populations (Mislove et al., 2011; Zickuhr, 2013)
  - Age (Longley et al., 2015), binned by ages 10-17, 18-29, 30-49, 50-64, and 65+ (Zickuhr, 2013)
  - Urban and rural populations (Hecht and Stephens, 2014; Zickuhr, 2013)
  - Median income (Zickuhr, 2013)
  - For all of these except median income, stabilize variance with a log transformation and add-one smoothing.

Model specification

- Following previous literature, we include covariates for:
  - Black population, Asian population, Latino/Hispanic populations (Mislove et al., 2011; Zickuhr, 2013)
  - Age (Longley et al., 2015), binned by ages 10-17, 18-29, 30-49, 50-64, and 65+ (Zickuhr, 2013)
  - Urban and rural populations (Hecht and Stephens, 2014; Zickuhr, 2013)
  - Median income (Zickuhr, 2013)
  - For all of these except median income, stabilize variance with a log transformation and add-one smoothing.
- We introduce terms for
  - Northern/eastern effect (demeaned latitudes and longitudes of block group centroids)
  - Coastal effect (squared terms for latitude and longitude)
- No analysis of sex as in Longley et al. (2015); Zickuhr (2013); Mislove et al. (2011), as those use name-based inference or survey data, but recall that sex is randomly distributed across block groups.

Motivation

1. All maps of geotagged tweets look like maps of population density
2. LOTS of studies use geotagged tweets
3. Need for large-scale, multivariate, statistical work

Data

1. Geotags in tweets
2. US Census
3. Mobile users

Model

1. A statistical test for random distribution over population
2. Validating the null model
3. Spatial autocorrelation and spatial errors
4. Model specification

Results

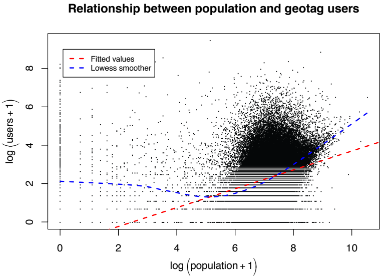
1. Test of null hypothesis
2. Spatial autocorrelation
3. Spatial errors model results

Discussion

1. Limitations
2. Take-away
3. Future work

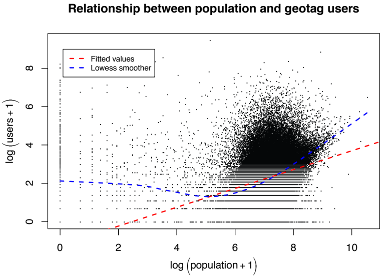
Test of null hypothesis

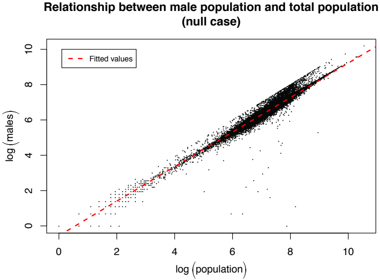
Ignoring spatial autocorrelation, and testing $H_0 : \beta_1 = 0$ with OLS, we get slope $\hat{\beta}_1 = .4916$ (.002996) and intercept $\hat{\beta}_0 = -1.219$ (.02143).



Test of null hypothesis

Ignoring spatial autocorrelation, and testing $H_0 : \beta_1 = 0$ with OLS, we get slope $\hat{\beta}_1 = .4916$ (.002996) and intercept $\hat{\beta}_0 = -1.219$ (.02143).

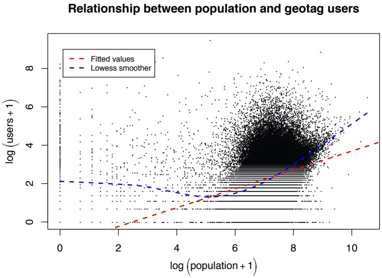


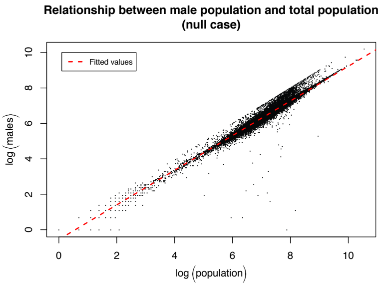


1 is far outside the 95% confidence interval for β_1 of $[0.4857, 0.4975]$. Same for only nonzero block groups, and >5, >10 tweet filters. No need to worry about spatial autocorrelation; reject the null!

Test of null hypothesis

Ignoring spatial autocorrelation, and testing $H_0 : \beta_1 = 0$ with OLS, we get slope $\hat{\beta}_1 = .4916$ (.002996) and intercept $\hat{\beta}_0 = -1.219$ (.02143).





1 is far outside the 95% confidence interval for β_1 of $[0.4857, 0.4975]$. Same for only nonzero block groups, and >5, >10 tweet filters. No need to worry about spatial autocorrelation; reject the null!

(Note that we no longer interpret $\hat{\alpha} = e^{\hat{\beta}_0}$).

Form of spatial autocorrelation

Values of Moran's I in the bivariate regression:

	Population vs Users	Population vs Male
2nn	.3699	.2336
4nn	.3550	.2142
6nn	.3398	.1996
8nn	.3270	.1883
Rook	.4166 (b)	.2125 (b)
	.3992 (rn)	.2201 (rn)
Queen	.4151 (b)	.2097 (b)
	.3919 (rn)	.2154 (rn)

For the Rook contiguity case and the Queen contiguity case, binary (b) and row-normalized (rw) weights give different values. (Nearly identical results for different filter levels).

Form of spatial autocorrelation

Values of Moran's I in the bivariate regression:

	Population vs Users	Population vs Male
2nn	.3699	.2336
4nn	.3550	.2142
6nn	.3398	.1996
8nn	.3270	.1883
Rook	.4166 (b)	.2125 (b)
	.3992 (rn)	.2201 (rn)
Queen	.4151 (b)	.2097 (b)
	.3919 (rn)	.2154 (rn)

For the Rook contiguity case and the Queen contiguity case, binary (b) and row-normalized (rw) weights give different values. (Nearly identical results for different filter levels).

All weights pick up spatial autocorrelation. Cholesky decomposition requires a symmetric matrix, so use binary Rook contiguity.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)

Note: *p<.0001

- No more spatial autocorrelation in residuals

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)

Note: *p<.0001

- No more spatial autocorrelation in residuals
- After controlling for other factors, population loses its significance (not so with OLS).

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- No more spatial autocorrelation in residuals
- After controlling for other factors, population loses its significance (not so with OLS).
- Area is significant, and +1% area $\implies$ +15.56% geotag users. Size overcomes the effects of population density (perhaps effect of tweeting on highways?)

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)

Note: *p<.0001

- No more spatial autocorrelation in residuals
- After controlling for other factors, population loses its significance (not so with OLS).
- Area is significant, and +1% area $\implies$ +15.56% geotag users. Size overcomes the effects of population density (perhaps effect of tweeting on highways?)
- +1% Hispanic/Latino population $\implies$ +1.533% geotag users. Direction, but not size, consistent with survey results (Zickuhr, 2013).

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- No more spatial autocorrelation in residuals
- After controlling for other factors, population loses its significance (not so with OLS).
- Area is significant, and +1% area $\implies$ +15.56% geotag users. Size overcomes the effects of population density (perhaps effect of tweeting on highways?)
- +1% Hispanic/Latino population $\implies$ +1.533% geotag users. Direction, but not size, consistent with survey results (Zickuhr, 2013).
- +1% Asian population $\implies$ +11.12% geotag users.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- No more spatial autocorrelation in residuals
- After controlling for other factors, population loses its significance (not so with OLS).
- Area is significant, and +1% area $\implies$ +15.56% geotag users. Size overcomes the effects of population density (perhaps effect of tweeting on highways?)
- +1% Hispanic/Latino population $\implies$ +1.533% geotag users. Direction, but not size, consistent with survey results (Zickuhr, 2013).
- +1% Asian population $\implies$ +11.12% geotag users.
- +1% Black population $\implies$ +4.29% geotag users. Contradicts survey results, but qualitative research shows active community on Twitter (Clark, 2014; Florini, 2014; Sharma, 2013).

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- No more spatial autocorrelation in residuals
- After controlling for other factors, population loses its significance (not so with OLS).
- Area is significant, and +1% area $\implies$ +15.56% geotag users. Size overcomes the effects of population density (perhaps effect of tweeting on highways?)
- +1% Hispanic/Latino population $\implies$ +1.533% geotag users. Direction, but not size, consistent with survey results (Zickuhr, 2013).
- +1% Asian population $\implies$ +11.12% geotag users.
- +1% Black population $\implies$ +4.29% geotag users. Contradicts survey results, but qualitative research shows active community on Twitter (Clark, 2014; Florini, 2014; Sharma, 2013).
- The latitude, both in linear and quadratic effects, is not significant.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- No more spatial autocorrelation in residuals
- After controlling for other factors, population loses its significance (not so with OLS).
- Area is significant, and +1% area $\implies$ +15.56% geotag users. Size overcomes the effects of population density (perhaps effect of tweeting on highways?)
- +1% Hispanic/Latino population $\implies$ +1.533% geotag users. Direction, but not size, consistent with survey results (Zickuhr, 2013).
- +1% Asian population $\implies$ +11.12% geotag users.
- +1% Black population $\implies$ +4.29% geotag users. Contradicts survey results, but qualitative research shows active community on Twitter (Clark, 2014; Florini, 2014; Sharma, 2013).
- The latitude, both in linear and quadratic effects, is not significant.
- Longitude is significant in both effects: block groups further east having more geotag users, and second block groups on both the east and west coasts have more geotag users.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages $\geq$65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income $\implies$ +1.66% geotag users.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
$long^2$	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income $\implies$ +1.66% geotag users.
- Tried to test for median income squared but got computational singularity; visual inspection of plot showed no evidence of nonlinear relationship, and linear effect is weak.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
$long^2$	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income ⟹ +1.66% geotag users.
- Tried to test for median income squared but got computational singularity; visual inspection of plot showed no evidence of nonlinear relationship, and linear effect is weak.
- +1% rural population ⟹ -5.72% geotag users, consistent with Hecht and Stephens (2014).

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income $\implies$ +1.66% geotag users.
- Tried to test for median income squared but got computational singularity; visual inspection of plot showed no evidence of nonlinear relationship, and linear effect is weak.
- +1% rural population $\implies$ -5.72% geotag users, consistent with Hecht and Stephens (2014).
- +1% 18-29 year olds $\implies$ +39.16% geotag users. Consistent with survey results.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income $\implies$ +1.66% geotag users.
- Tried to test for median income squared but got computational singularity; visual inspection of plot showed no evidence of nonlinear relationship, and linear effect is weak.
- +1% rural population $\implies$ -5.72% geotag users, consistent with Hecht and Stephens (2014).
- +1% 18-29 year olds $\implies$ +39.16% geotag users. Consistent with survey results.
- +1% 50-64 year olds $\implies$ -17.93% geotag users. Consistent with survey results.

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
$long^2$	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income $\implies$ +1.66% geotag users.
- Tried to test for median income squared but got computational singularity; visual inspection of plot showed no evidence of nonlinear relationship, and linear effect is weak.
- +1% rural population $\implies$ -5.72% geotag users, consistent with Hecht and Stephens (2014).
- +1% 18-29 year olds $\implies$ +39.16% geotag users. Consistent with survey results.
- +1% 50-64 year olds $\implies$ -17.93% geotag users. Consistent with survey results.
- Teenage population predicts fewer geotag users (than excluded group, ages <10).

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
$long^2$	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages ≥65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income $\implies$ +1.66% geotag users.
- Tried to test for median income squared but got computational singularity; visual inspection of plot showed no evidence of nonlinear relationship, and linear effect is weak.
- +1% rural population $\implies$ -5.72% geotag users, consistent with Hecht and Stephens (2014).
- +1% 18-29 year olds $\implies$ +39.16% geotag users. Consistent with survey results.
- +1% 50-64 year olds $\implies$ -17.93% geotag users. Consistent with survey results.
- Teenage population predicts fewer geotag users (than excluded group, ages <10).
- Elderly population predicts more geotag users (than excluded group, ages <10).

Spatial errors model results

	Dependent variable:	
	log(user+1)	s.e.
log(population+1)	-.01218	(.008081)
log(area)	.1556*	(.001760)
log(hispanic+1)	.01533*	(.002066)
log(asian+1)	.1112*	(.001576)
log(black+1)	.04292*	(.001576)
lat (demeaned)	-.006992	(.0007052)
lat^2	-1.641-e5	(9.505-e5)
long (demeaned)	.02306*	(.0002739)
long^2	8.777-e5*	(1.411-e5)
med income ($10K)	.01661*	(.0006857)
log(rural+1)	-.05722*	(.001096)
log(ages 10-17+1)	-.09831*	(.003712)
log(ages 18-29+1)	.3916*	(.004423)
log(ages 30-49+1)	.06362*	(.006731)
log(ages 50-64+1)	-.1793*	(.006953)
log(ages $\geq$65+1)	.09675*	(.003940)
Intercept	1.3382*	(.1916)
Note:	*p<.0001	

- +$10K median income $\implies$ +1.66% geotag users.
- Tried to test for median income squared but got computational singularity; visual inspection of plot showed no evidence of nonlinear relationship, and linear effect is weak.
- +1% rural population $\implies$ -5.72% geotag users, consistent with Hecht and Stephens (2014).
- +1% 18-29 year olds $\implies$ +39.16% geotag users. Consistent with survey results.
- +1% 50-64 year olds $\implies$ -17.93% geotag users. Consistent with survey results.
- Teenage population predicts fewer geotag users (than excluded group, ages $<$10).
- Elderly population predicts more geotag users (than excluded group, ages $<$10).
- (Intercept not meaningful for log dependent variable.)

Motivation

1. All maps of geotagged tweets look like maps of population density
2. LOTS of studies use geotagged tweets
3. Need for large-scale, multivariate, statistical work

Data

1. Geotags in tweets
2. US Census
3. Mobile users

Model

1. A statistical test for random distribution over population
2. Validating the null model
3. Spatial autocorrelation and spatial errors
4. Model specification

Results

1. Test of null hypothesis
2. Spatial autocorrelation
3. Spatial errors model results

Discussion

1. Limitations
2. Take-away
3. Future work

Limitations

- Doing things at scale gives lots of statistical power, but we lose the ability to do meaningful visual diagnostics and find interesting outliers
- We corrected for one type of misspecification, spatial autocorrelation, but there is potentially unexplored structure in the residuals
- Are other relevant models for dependent/spatial data, such as disease mapping, conditional autoregressive (CAR) models, Gaussian Process regression...
- 'Plurality' placement yields a few hundred block groups with more users than population; need better way to uniquely locate
- Don't account for foreign tourists (in 2013, 1 tourist for every 4.5 people in US)
- Still need assumption that demographics of a block group represent the geotag users there
- Add-one smoothing can produce some artifacts
- We don't look at time; we use data from 2010 but tweets from 2013, and geotag trends are almost certainly not stationary
- This is only for the US; doesn't generalize to any other country

Take-away

What are our methodological recommendations?

Take-away

What are our methodological recommendations?

- Geotagged tweets are *not* representative. Which means we can't use them to make valid inferences about any larger population.

Take-away

What are our methodological recommendations?

- Geotagged tweets are *not* representative. Which means we can't use them to make valid inferences about any larger population.
- Researchers don't need to stop using geotagged tweets, but we need to stop assuming that results generalize to larger populations and using this assumption as our driving motivation.

Take-away

What are our methodological recommendations?

- Geotagged tweets are *not* representative. Which means we can't use them to make valid inferences about any larger population.
- Researchers don't need to stop using geotagged tweets, but we need to stop assuming that results generalize to larger populations and using this assumption as our driving motivation.
- If we want generalizability, the easiest way to probably to directly demonstrate that we can infer a given external trend from tweets (comes with its own problems...).

Take-away

What are our methodological recommendations?

- Geotagged tweets are *not* representative. Which means we can't use them to make valid inferences about any larger population.
- Researchers don't need to stop using geotagged tweets, but we need to stop assuming that results generalize to larger populations and using this assumption as our driving motivation.
- If we want generalizability, the easiest way to probably to directly demonstrate that we can infer a given external trend from tweets (comes with its own problems...).
- We can (and should) still study geotagged tweets, Twitter users, and Twitter...

Take-away

What are our methodological recommendations?

- Geotagged tweets are *not* representative. Which means we can't use them to make valid inferences about any larger population.
- Researchers don't need to stop using geotagged tweets, but we need to stop assuming that results generalize to larger populations and using this assumption as our driving motivation.
- If we want generalizability, the easiest way to probably to directly demonstrate that we can infer a given external trend from tweets (comes with its own problems...).
- We can (and should) still study geotagged tweets, Twitter users, and Twitter... as inherently interesting social and cultural phenomena.

Take-away

What are our methodological recommendations?

- Geotagged tweets are *not* representative. Which means we can't use them to make valid inferences about any larger population.
- Researchers don't need to stop using geotagged tweets, but we need to stop assuming that results generalize to larger populations and using this assumption as our driving motivation.
- If we want generalizability, the easiest way to probably to directly demonstrate that we can infer a given external trend from tweets (comes with its own problems...).
- We can (and should) still study geotagged tweets, Twitter users, and Twitter... as inherently interesting social and cultural phenomena.
- We need to take statistical approaches, and many relevant models already exist.

Future work

- Re-test with 2013 ACS 1-year estimates: incomplete and lower resolution but more current
- Repeat analysis in other countries
- Filter out ‘non-personal’ users (Guo and Chen, 2014)
- Filter out foreign tourists by collecting all geotagged tweets or profile info
- Find better ways to uniquely place users
- Model demographic differences in different levels of usage
- Apply other spatial models
- Long-term spatio-temporal modeling
- With knowledge of demographic biases, we could model demographic changes

Thank you! Questions?

Momin M. Malik <momin.malik@cs.cmu.edu>
Hemank Lamba <hemank.lamba@cs.cmu.edu>
Constantine Nakos <cnakos@andrew.cmu.edu>
Jürgen Pfeffer <jpfeffer@cs.cmu.edu>

Slides at http://mominmalik.com/Malik_ICWSM2015_slides.pdf

References

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3):247–267.

Anselin, L. and Rey, S. (1991). Properties of tests for spatial dependence in linear regression models. *Geographical Analysis*, 23(2):112–131.

Anselin, L., Sridharan, S., and Gholston, S. (2007). Using exploratory spatial data analysis to leverage social indicator databases: The discovery of interesting patterns. *Social Indicators Research*, 82(2):287–309.

Bivand, R., Hauke, J., and Kossowski, T. (2013a). Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis*, 45(2):150–179.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63(18):1–36.

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013b). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090.

Clark, M. D. (2014). *To tweet our own cause: A mixed-methods study of the online phenomenon "Black Twitter"*. PhD thesis, The University of North Carolina at Chapel Hill, School of Journalism and Mass Communication.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., and Zook, M. (2013). Beyond the geotag: Situating "big data" and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2):130–139.

Doran, D., Gokhale, S., and Dagnino, A. (2013). Human sensing for smart cities. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1323–1330.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287.

Florini, S. (2014). Tweets, tweeps, and signifyin': Communication and cultural performance on "Black Twitter". *Television & New Media*, 15(3):223–237.

Frias-Martinez, V., Soto, V., Hohwald, H., and Frias-Martinez, E. (2012). Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, PASSAT/SocialCom '12, pages 239–248.

Gaetan, C. and Guyon, X. (2012). *Spatial Statistics and Modeling*. Springer Series in Statistics. Springer.

Ghosh, D. D. and Guha, R. (2013). What are we 'tweeting' about obesity?: Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, 40(2):90–102.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you?: Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

Guo, D. and Chen, C. (2014). Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS*, 18(3):370–384.

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246.

Hecht, B. and Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM '14, pages 197–205.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. (2012). Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 769–778.

Kamath, K. Y., Caverlee, J., Lee, K., and Cheng, Z. (2013). Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 667–678.

Kinsella, S., Murdock, V., and O'Hare, N. (2011). "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 61–68.

Kumar, S., Hu, X., and Liu, H. (2014). A behavior analytics approach to identifying tweets from crisis regions. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 255–260.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5).

Lin, Y.-R. and Margolin, D. (2014). The ripple of fear, sympathy and solidarity during the Boston bombings. *EPJ Data Science*, 3(1).

Liu, Y., Kliman-Silver, C., and Mislove, A. (2014). The tweets they are a-changinÕ: Evolution of Twitter users and behavior. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM '14.

Longley, P. A., Adnan, M., and Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, 47(2):465–484.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. (2011). Understanding the demographics of Twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 554–557.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5):e64417.

Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J., and Liu, H. (2014). Finding eyewitness tweets during crises. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, ACL LACSS '14, pages 23–27.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. (2013). Is the sample good enough?: Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM '13.

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., and Brownstein, S. J. (2014). A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res*, 16(10):e236.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

Sharma, S. (2013). Black Twitter?: Racial hashtags, networks and contagion. *New Formations: A Journal of Culture/Theory/Politics*, 78(1).

Shelton, T., Poorthuis, A., Graham, M., and Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum*, 52(0):167 – 179.

Stephens, M. and Poorthuis, A. (2014). Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems*.

Sylvester, J., Healey, J., Wang, C., and Rand, W. M. (2014). Space, time, and hurricanes: Investigating the spatiotemporal relationship among social media use, donations, and disasters. Technical Report Research Paper No. RHS 2441314, Robert H. Smith School.

Takhteyev, Y., Gruzd, A., and Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, 34(1):73–81.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM '14, pages 505–514.

van Dijck, J. (2013). Chapter 4: Twitter and the paradox of following and trending. In *The Culture of Connectivity: A Critical History of Social Media*, pages 68–88. Oxford University Press.

van Liere, D. (2010). How far does a tweet travel?: Information brokers in the Twitterverse. In *Proceedings of the International Workshop on Modeling Social Media*, MSM '10, pages 6:1–6:4.

Wang, D., Al-Rubaie, A., Davies, J., and Clarke, S. (2014). Real time road traffic monitoring alert based on incremental learning from tweets. In *2014 IEEE Symposium on Evolving and Autonomous Learning Systems*, EALS '14, pages 50–57.

Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. (2013). Who, where, when and what: Discover spatio-temporal topics for Twitter users. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 605–613.

Zickuhr, K. (2013). Location-base services. Technical Report Pew Internet and American Life Project, Pew Research Center.