Using Twitter Data for Environmental Migration Research:

An Example of Migration and Perceptions of Climate Change in Alaska

Guangqing Chi

The Pennsylvania State University

Presentation to the International Conference of
Migration, Environment and Climate: What Risks Inequalities?
Paris, France, October 22-23, 2018

Acknowledgements

NSF SES-1823633: Junjun Yin, Eric Plutzer, Jennifer Van Hook, Heng Xu

NSF OPP-1745369: Ming Xiao

Donghui Wang

SOCIAL SCIENCE RESEARCH INSTITUTE
Supporting Novel Interdisciplinary Research to Address Critical Human and Social Problems

POPULATION RESEARCH INSTITUTE
A Community of Population Scientists at Penn State

Challenges of Environmental Migration Research

Migration data

- Data collection is slow, labor intensive, and expensive
- Surveys
  - Limited samples
  - Retrospective information are not accurate
- Census and other data collected by governments
  - Available at aggregated levels
  - US IRS data: county level; two-years release lag
  - US Census data: county level
- Data collection for migration (or mobility) due to extreme, sudden disasters cannot provide timely information for disaster rescue and emergency management

Big (Social Media) Data

- The potential of Big Data lies in the ability to collect massive amounts of information from a large group of individuals.
- The large N allows the potential for high resolution classification and the possibility of generating samples of individuals in small or hard-to-reach populations, assisting in the real-world study of various population dynamics.
- Big Data allow researchers to track changes in populations very quickly, in real time. Think about population movements in response to Hurricane Michael.

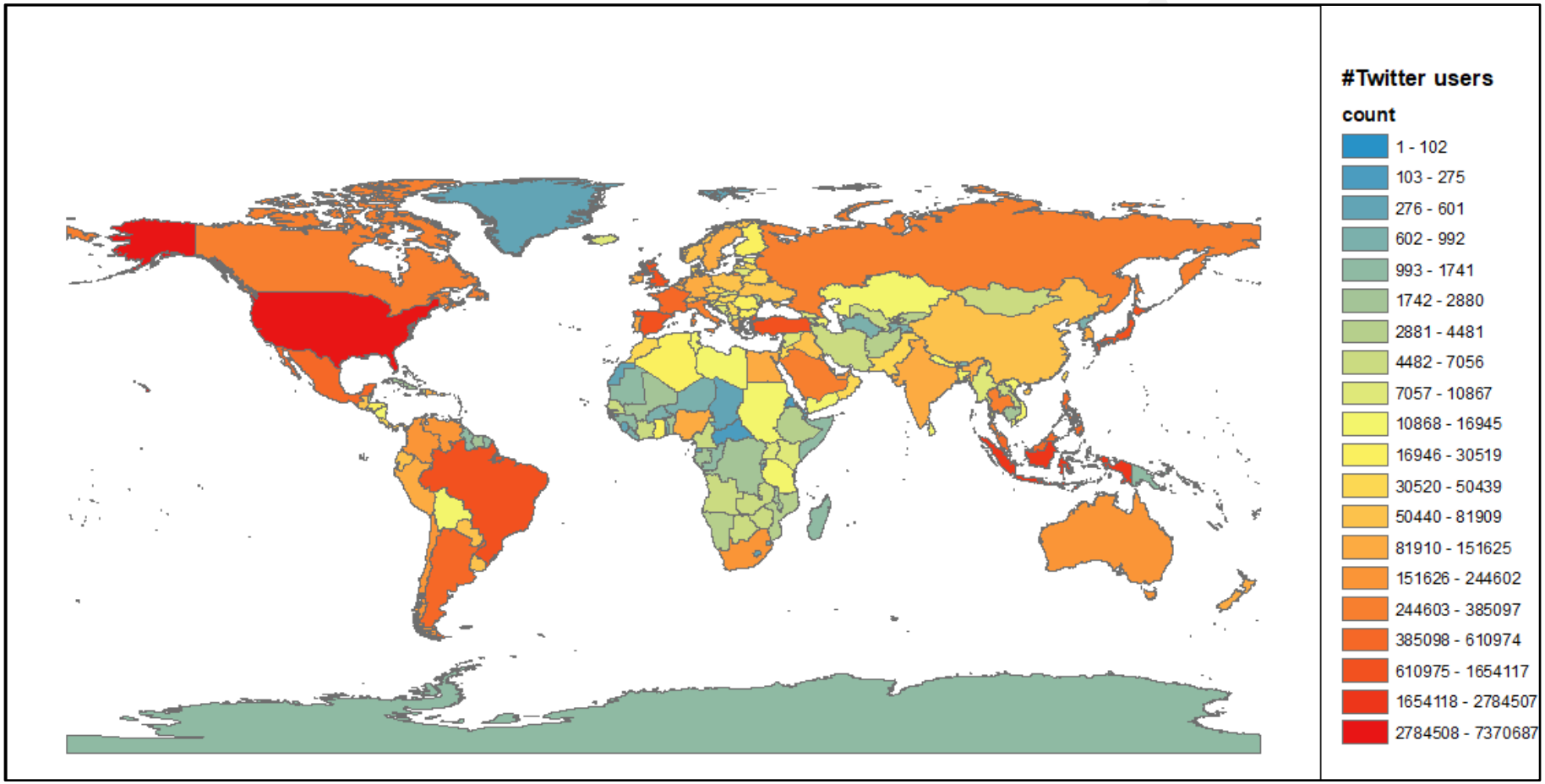
Twitter data vs. Other Big Data

- Twitter is arguably the most popular information source for the scientific research community due to its accessibility.
- The geolocations of tweets either are geotagged or could be inferred.
- Its contents can be mined for valuable information.

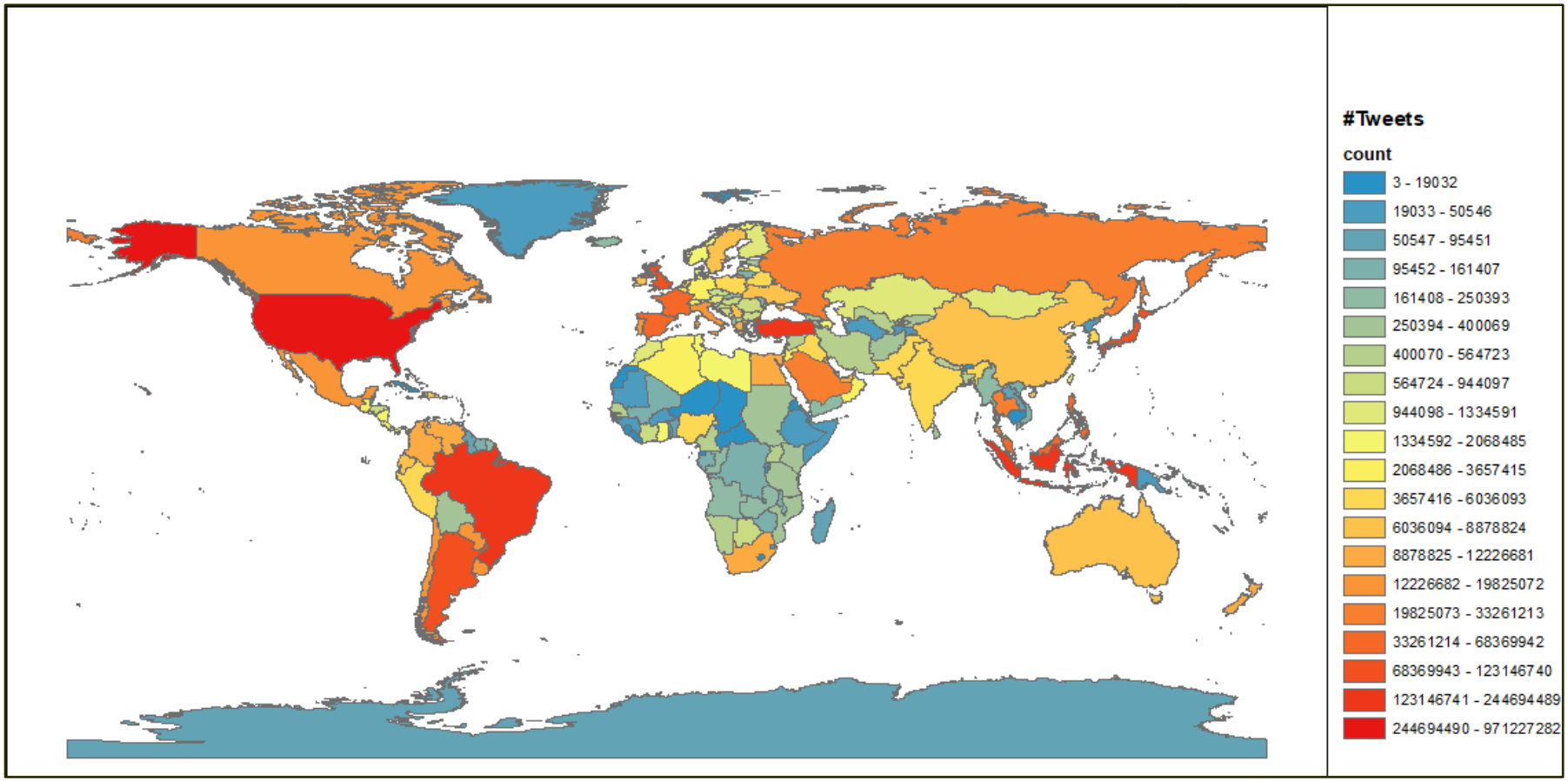
Information from Twitter

1. Profile (e.g., user ID, image, hometown, country, etc)
2. Followers
3. People to follow
4. Retweet
5. Like
6. Twitter message (sometimes including hashtags, pictures) → perceptions of climate change
7. Location and time of the message

Number of geo-located Twitter users, February - December 2014



Number of geo-located Tweets, February - December 2014



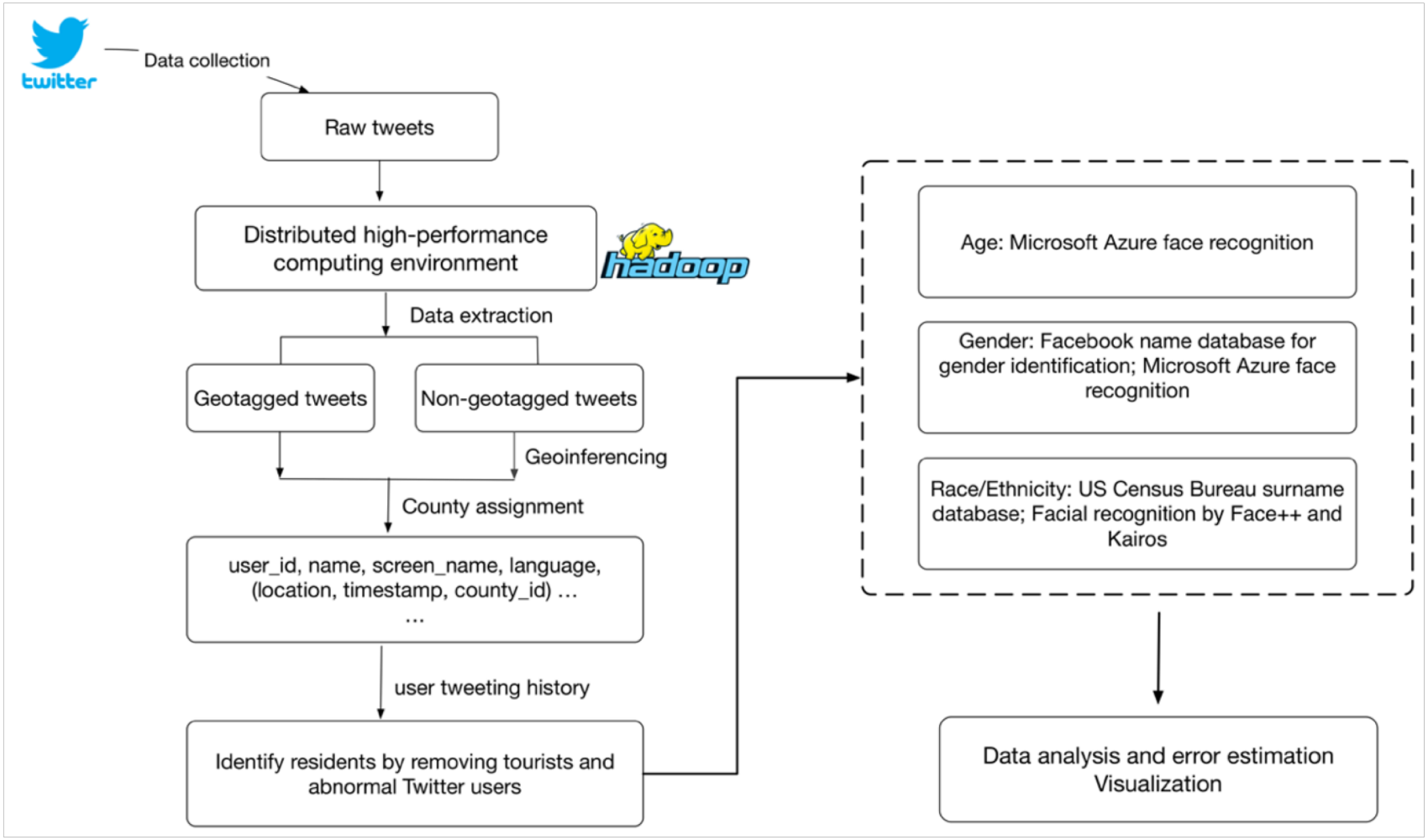
Twitter Data

- Challenge: Representativeness of the data
- Other concerns:
  - the hardships of dealing with slang, sarcasm, and unconventional forms of written expression, including hashtags, emoticons, and acronyms;
  - dealing with the fact that not all Twitter users are humans but bots, which can distort the results; and
  - managing the cost of obtaining, storing, and cleaning the massive datasets from Twitter.

NSF # SES-1823633:
The Generalizability and Replicability of Twitter Data for Population Research

- Aim 1: evaluating the extent to which Twitter users (mis)represent the population across different demographic groups.
- Aim 2: testing the feasibility of developing weights that, when applied to Twitter data, make the results more representative of the underlying population.
- Aim 3: testing the feasibility of using Twitter data to estimate migration at the county level by comparing to the IRS migration data, and estimate Puerto Rico migrants to the US continent after Hurricane Maria.

Overall process of estimating demographics of Twitter users



Biases of Twitter User Estimates at the County Level in 2014

	Population Median	Twitter Users Median	Median Percentage Error	Median Absolute Percentage Error
Female	50.4%	52.3%	4.2%	8.3%
Male	49.6%	47.7%	–4.2%	8.3%
Age 20–24	10.1%	28.5%	173.7%	173.7%
Age 25–34	19.1%	49.3%	155.6%	155.6%
Age 35–44	18.3%	14.3%	–20.7%	26.0%
Age 45–54	19.8%	5.2%	–74.5%	74.5%
Age 55–64	18.7%	1.9%	–90.1%	90.1%
Age 65+	22.5%	1.4%	–93.7%	93.7%
Non-Hispanic White	89.2%	72.6%	–15.9%	17.6%
Non-Hispanic Black	2.3%	12.6%	369.4%	369.4%
Hispanics	4.1%	5.9%	31.6%	45.7%

Methods of Weighting adjustment

1. Simple ratio weights
2. Raking extension
3. Propensity scores
4. Matching methods
5. Multilevel regression with post-stratification

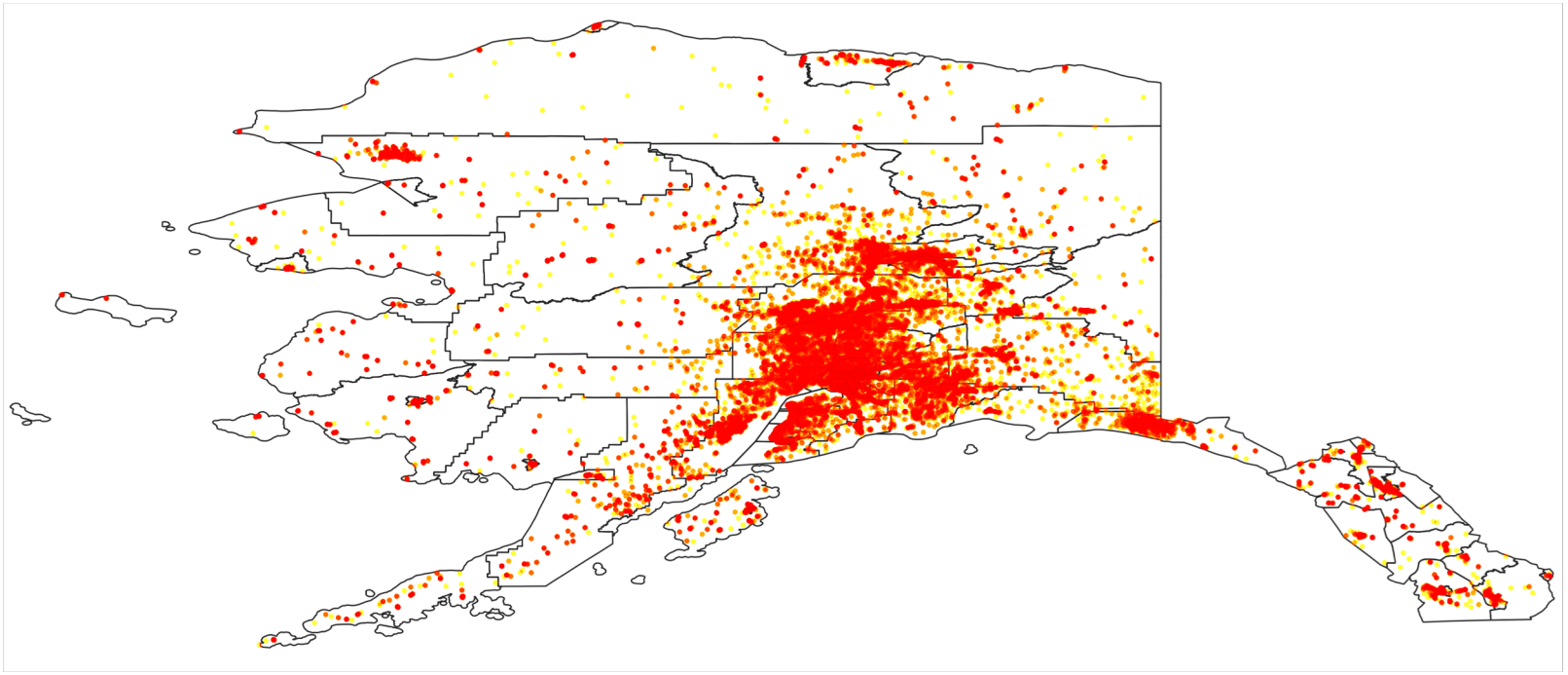
Split-Sample Design

1. Training data: a random selection of counties from 2014-2017; Test data: the rest counties
   - assess the representativeness of Twitter data,
   - model the bias,
   - create weights,
   - apply the weights to the test data, and
   - check the difference between weighted Twitter population estimates and the Census estimates.

Split-Sample Design

2. Train the 2014-2015 data and test 2016-2017 data
   - Use the 2014-2015 data to assess the representativeness of Twitter data relative to Census data for all U.S. counties,
   - Model the bias as a function of county characteristics (e.g., proportions of demographic groups, rurality, region of the county, telecommunication infrastructure),
   - Create weights for 2014-2015 based on these analyses, and
   - Apply weights to 2016-2017 Twitter data and check how different the weighted Twitter population estimates are from the Census estimates of 2016-2017.

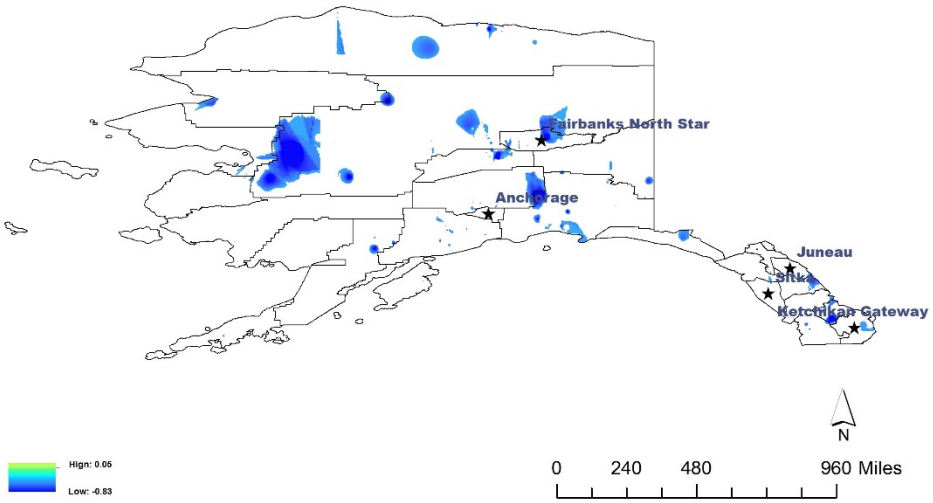
Migration and Perceptions of Climate Change in Alaska



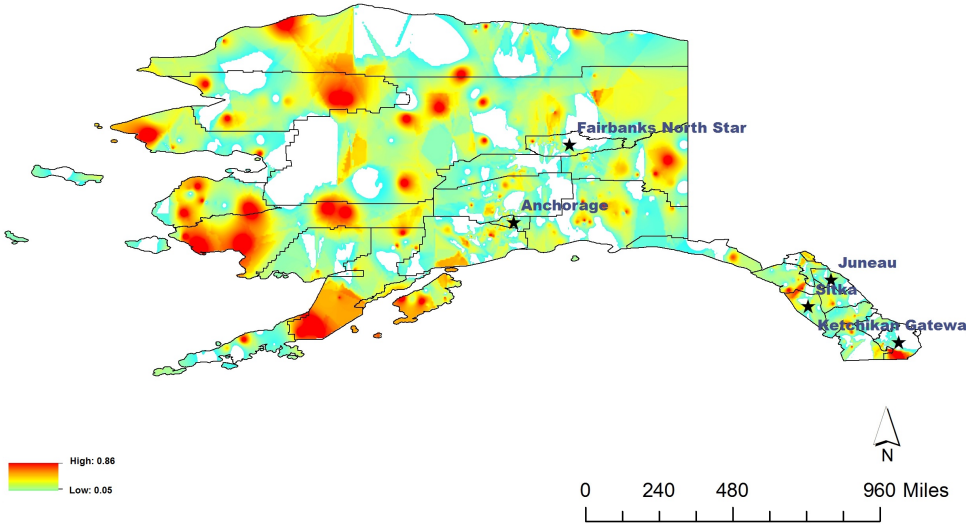
The Geo-tagged Tweets in Alaska, 2014

Sentiments toward climate change based on geo-tagged weets in Alaska, 2014

Negative

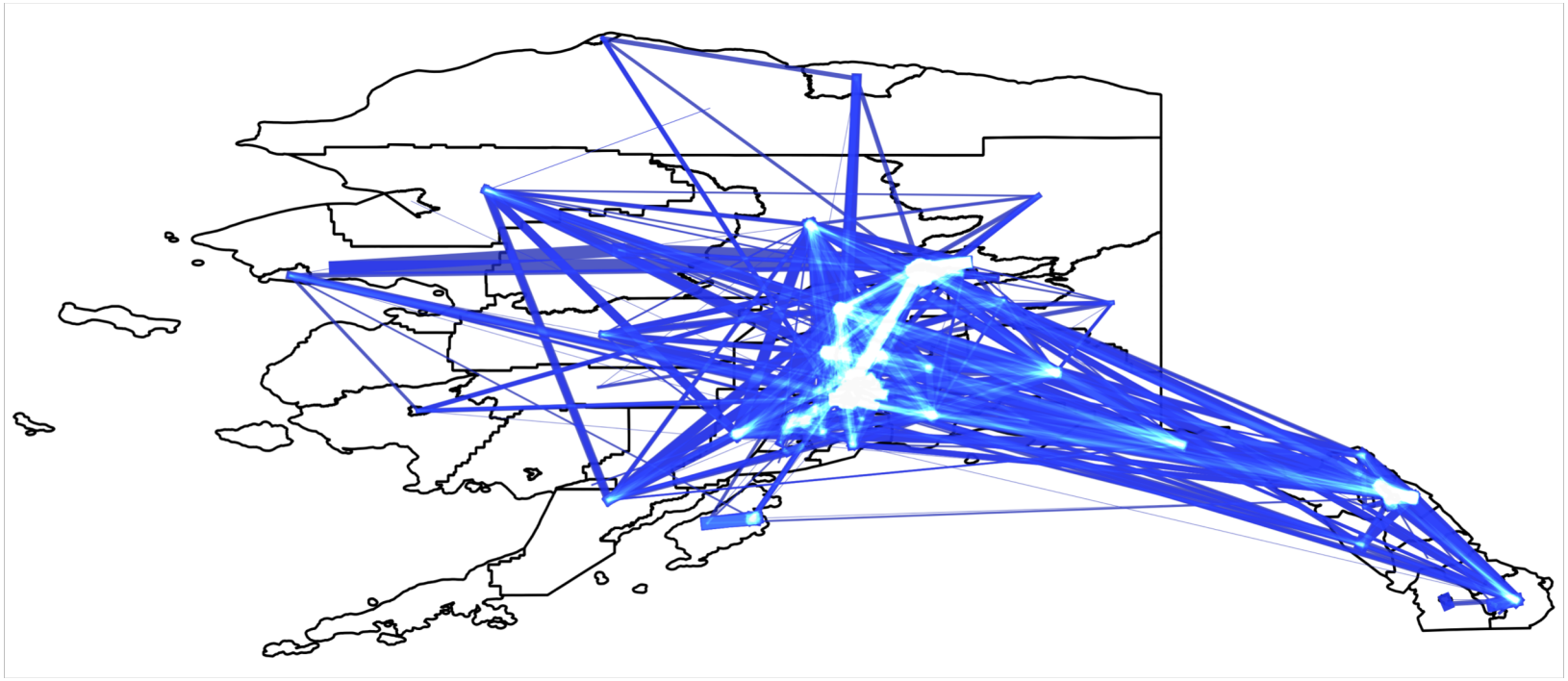


Positive



Keywords used: climate, weather, temperature, flood(flooding), rain, snow, ice, freeze, thaw, global warming, atmosphere, glacier, coastal erosion, degradation, landscape, damage, catastrophe, crisis, risk, vulnerability, endanger, hazard

Flows of Twitter users at the census tract level, Alaska, 2014



Movement flows Twitter users in Alaska

Take-home messages

1. Twitter data could be a great resource for environmental migration research, especially for time-sensitive research
2. Lots of work need to be done in order to produce robust and generalizable results