How we can use Twitter data to better understand weather-related depression.

The weather can have a profound influence on many people's emotional states, but until now it has been difficult to quantify these effects. The advent of social media has now made it much easier to explore the relationship between seasonality and the prevalence of depression. In new research which uses data from more than 600 million tweets over a one year period, **Wei Yang**, *along with* **Lan Mu** *and* **Ye Shen** *find that climate risk factors for depression are different and localized, depending on the area in question. They write that using such social media data has benefits over traditional data collecting methods, and may have the potential to transform clinical practice for some diseases.*

Depression has a high prevalence in the US, with 6.7 percent of adults over 18 reporting having had at least one depressive episode in the past year in 2013. Research has shown that climate impacts are related to this common chronic stress-related disorder or with negative emotions. However, previous studies have yielded mixed results due to limitations such as small and maybe biased datasets, inconsistent spatial and temporal data domains and lags between data collection and analysis. In recent years, social media has received considerable attention as a new data source for health research, as it can provide an enormous stream of data about people's lives and behavior. In new research, we explore the relationship between seasonality and the prevalence of depression using social media data.

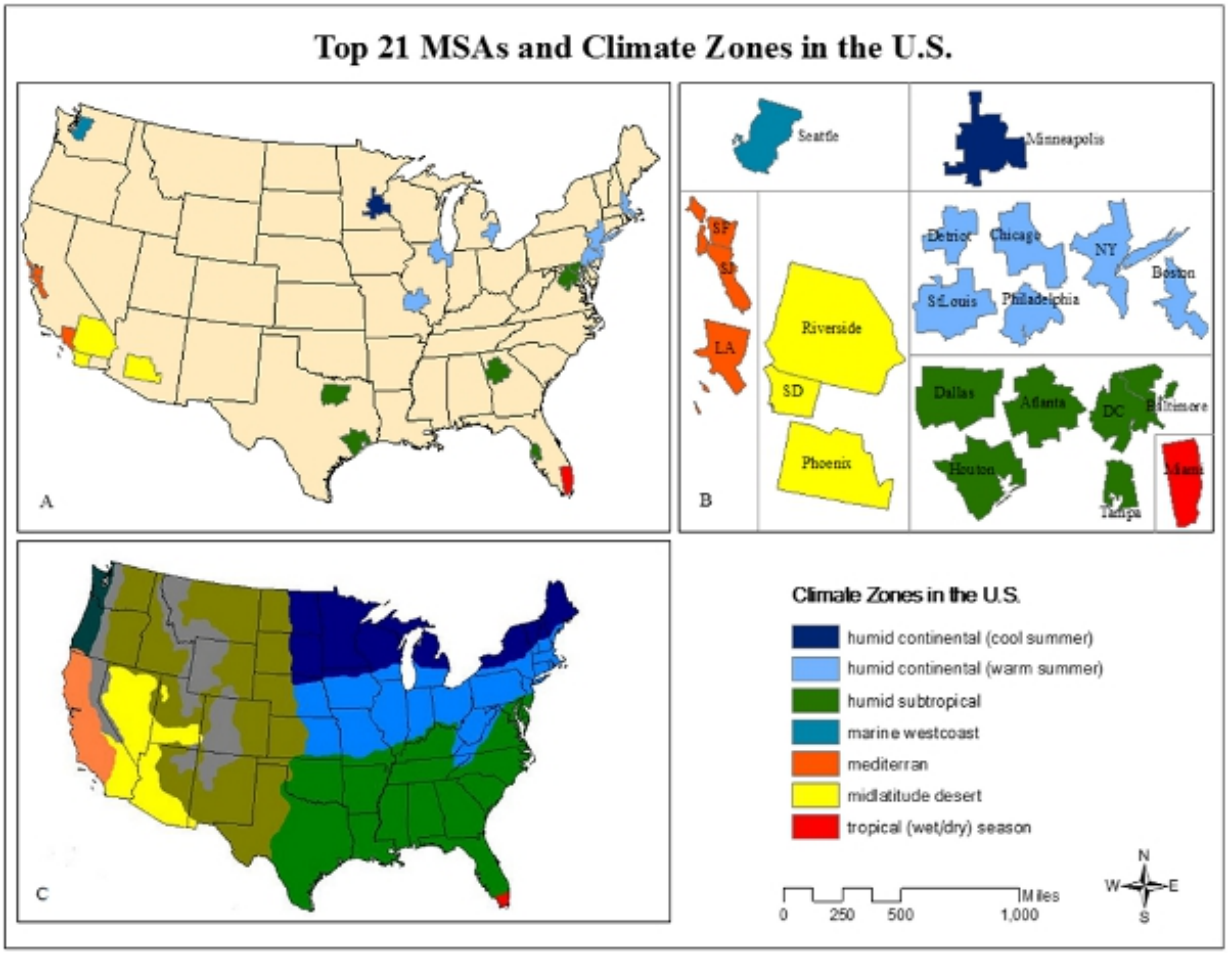
We investigated the interaction between the rate of tweets expressing depressed feelings, climate, seasonality, and geographical locations by exploring Twitter data from textual, spatial and temporal aspects in the US. We downloaded more than 600 million tweets from September 2013 to September 2014 using Twitter Steaming API. We only kept tweets written in English and posted in the US. We first used keywords depress or its variations to significantly reduce the size of our data. We also adopted an advanced text mining algorithm – non-negative matrix factorization (NMF) – to differentiate the word context related to true depressed feelings from those that are not talking about depression.

For geographic analysis, the study areas are the top 20 and the 34th metropolitan statistical areas (MSAs) in the US. The 34th MSA, San Jose is included because it is a fast growing MSA at the heart of Silicon Valley, and the headquarters of the hi-tech industry in the US. In order to capture the temporal variation of climate in different US locations, we divided the 21 MSAs into different groups by climate zones (Table 1). Figure 1A shows the exact location for each MSA. Figure 1C shows the classification of climate zones in the US. For better visualization, we created a conceptual compact map (Figure 1B). Each box represents a climate zone. To optimize visualization, we used varied map scales for the MSAs and only kept the shapes and relative positions within each climate zone.

Table 1 – Climate zones of the continental US

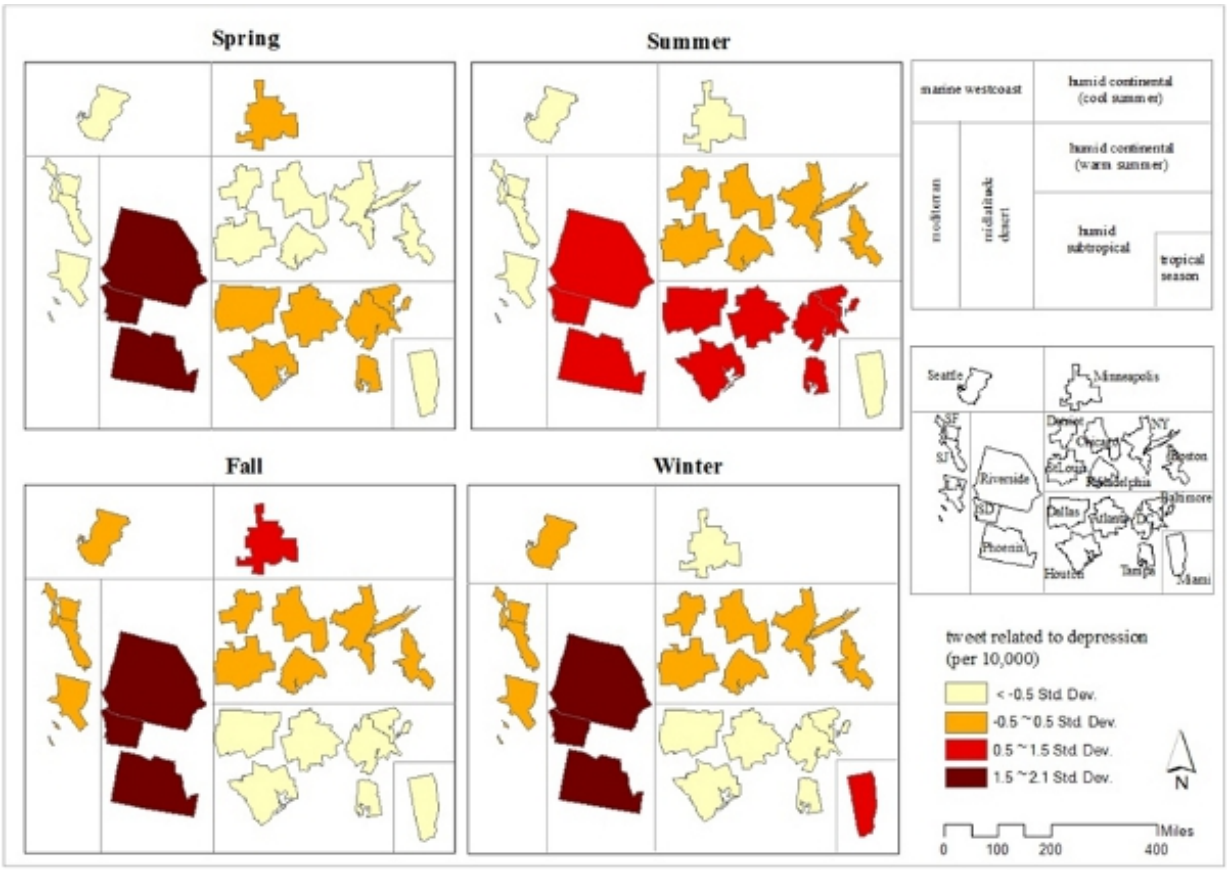
Climate	MSA short name
Humid continental (cool summer)	Minneapolis
Humid continental (warm summer)	St.Louis, Philadelphia, NewYork, Chicago, Detroit, Boston
Humid subtropical	DC, Baltimore, Dallas, Atlanta, Tampa, Houston
Marine westcoast	Seattle
Mediterranean	LA, San Jose, SF
Midlatitude desert	Riverside, San Diago, Phoenix
Tropical (wet/dry) season	Miami

Figure 1 – Climate zones of the top 21 MSAs in the US



To detect the seasonality effect on the spatiotemporal pattern of depression, we calculated the depression rates of the four seasons separately for each of the climate zones. The gradient from red into yellow represents depression rate from high to low (Figure 2). We observed that in the summer, on the east half of the US, the depression rates of different climate zones from the highest to the lowest are: humid subtropical, humid continental (warm summer), and humid continental (cool summer). But the sequence is reversed in the fall. Nationwide, we found that the mid-latitude desert climate zone always has the highest depression rate among all climate zones all year round.

Figure 2 – Rate of tweets related to depression in different climate zones and seasons



In order to further understand how the depression rate changes by season and what local climatic variables are most salient in explaining the spatiotemporal variations in depression rate in each climate zone, we conducted a regression model within each climate zone. The dependent variable is the ratio of tweets related to depression. The independent variables are relative humidity, temperature, sea level pressure, precipitation, snowfall, wind speed, globe solar radiation, and length of day. We found that the climatic risk factors for depression were different and localized. For example, wind speed plays a significant role in all climate zones. Sea level pressure and length of the day only matter in a single zone. Snowfall has a negative relationship with depression. Globe solar radiation has a positive relationship with depression. In the humid continental (warm summer) climate zone, relative humidity, precipitation and wind speed explain the most variation in depression rate. In the humid subtropical climate zone, temperature and globe solar radiation explain the most. In the Mediterranean and mid-latitude desert climate zones, relative humidity plays the most important role.

Our research is novel: in depression related research, the relationship between depression and climate has never been analyzed at the MSA aggregation level. Our use of social media date to study depression also avoids and solves the shortcomings of using traditional data collecting methods in health studies. Social media can protect users from exposing their identities face-to-face, thus this approach is better in terms of confidentiality and anonymity. Also, data acquired from social media can be seen as near-real time. Researchers do not have temporal or geographical constrains for collecting data, and it is faster and more cost-effective for analyzing health data. Another important improvement of using location-based social media data is that researchers do not need to manually geo-reference the data. This can help to avoid unnecessary errors.

Social media has become a phenomenon and is becoming a lasting resource for social science. However, social media data are not necessarily social science data. In our work, we have also proposed a framework to conceptualize the procedure of studying geographically distributed health issues using location-based social media data. This framework can help us understand how social and behavioral interventions influence humans' health and illness. Our framework implies that social media may have the potential to transform clinical practice considering some particular disease conditions. In addition, this framework can be used to detect major event outbreaks, such as flu and earthquakes.

This article is based on the paper, Yang, W., L. Mu & Y. Shen (2015) Effect of climate and seasonality on depressed mood among twitter users. Applied Geography, 63, 184-191.

Featured image credit: DJ Lein (Flickr, CC-BY-NC-SA-2.0)

Please read our comments policy before commenting.

Note: This article gives the views of the author, and not the position of USAPP – American Politics and Policy, nor the London School of Economics.

Shortened URL for this post: **http://bit.ly/1Z6OykH**

About the authors

Wei Yang – *University of Southern California*

Wei Yang is a Lecturer in Spatial Sciences with the Spatial Sciences Institute at the University of Southern California. She was a PhD candidate in the Department of Geography, the University of Georgia when writing this paper.

Lan Mu- *University of Georgia*

Lan Mu is Associate Professor of Geography at the University of Georgia. Her research interests include geographic information science (GIScience), GISceicne for health and the environment, and computational geometry.

Ye Shen – *University of Georgia*

Ye Shen is Assistant Professor of Biostatistics in the Department of Epidemiology and Biostatistics at the College of Public Health, the University of Georgia. His research interests include Longitudinal Data Analysis and Spatial Statistics.