

Track Spatiotemporal Spread of Public Concerns on Ebola in the U.S. via Twitter

Ying Liu, Feixiong Luo and Guofeng Cao

Department of Geosciences,
Texas Tech University
Email: {ying.liu, feixiong.luo, [guofeng.cao](mailto:guofeng.cao@ttu.edu)}@ttu.edu

1. Introduction

Emergence of social media has dramatically changed ways of people obtaining information and learning about the world. General public is increasingly dependent on online social media for news retrieval and information acquisition. Twitter, e.g., has become one of the most popular online social media for users to share information. With the continue advances of location-based services and wide spread of location-aware smart devices, the collection of tweets include rich amount of information, especially spatial and temporal information, regarding reactions of public on real-time events. Previous studies have successfully used Twitter data to track public reactions on vital natural and social events, e.g. fire hazard (Kent & Capello, 2013), earthquake (Sakaki et al., 2010), hurricane (Gupta et al., 2013), and epidemic disease (Signorini et al., 2011).

In March 2014 Ebola, a fulminating infectious and deadly disease, broke out in West Africa reported by the World Health Organization (WHO). Up to November 11, 2014, Centers for Disease Control and Prevention (CDC) posted that 14,413 Ebola cases were confirmed and Ebola had taken 5,504 people's lives. Effective methods have yet been developed to restrain the wide spread of Ebola over the world. Besides West African countries, Ebola also affects the United States (U.S.). On September 30, 2014 the CDC reported the first confirmed affected case of Ebola in the U.S. territory, a man who travelled to Dallas, Texas from Liberia. Just a few days later, two healthcare workers in Dallas and a medical aid worker who had just returned back to New York City from Guinea were diagnosed positive for Ebola.

Timely information regarding infectious diseases is critical to prevent a wider spread of the disease and public panic (Signorini et al., 2011). Therefore, a major objective of this study is to show capacities of Twitter to identify and track public concerns on time critical events, such as Ebola-related events in the United States. To fulfil the study objective, we first analyse the changes in the number of tweets regarding Ebola over thirty-eight days. Second, kernel density estimation (KDE) process is conducted to detect the occurring places of Ebola-related events. To adjust the impact of population on the total number of Ebola-related tweets, a population relative kernel density (PRKD) is developed as a new indicator to more reasonably explore the occurring places of Ebola cases. Finally, based on the spatio-temporal changing pattern of Ebola-related tweets, the potential of using Twitter data to track the spatiotemporal spread of public concerns over public health events is discussed.

2. Data and Methodology

Tweets for thirty-eight days (i.e., from September 25, 2014 to November 5, 2014) were crawled from Twitter Streaming API website. The crawled tweets were then filtered by the key words on Ebola (i.e., Ebola, EVD, and EHF). Finally, 49,949 tweets related with Ebola were totally extracted. The LandScan population dataset resolution was used in this study to evaluate the impact of population density on the number of tweets.

To explore the spatiotemporal characteristics of public concerns over Ebola in the U.S., a kernel density estimation (KDE) (Parzen, 1962) map showing probability densities of tweets with respect to Ebola was produced for each day between September 25, 2014 and November 5, 2014. The number of tweets regarding Ebola in a certain area is greatly affected by the region's population. It can be expected that a city with large population is likely to have relatively large number of tweets regarding Ebola even though no Ebola case occurs in the city. To mitigate this base population effect, we adjusted the density output of KDE by the population density (Equation 1), namely population relative kernel density (PRKD), to more reasonably explore the density distribution of Ebola-related tweets.

$$PRKD = \frac{KDE}{PopDen} \times 10^5 \quad (1)$$

where 10^5 is a scale factor to avoid too small PRKDs, and *PopDen* denotes population density that was obtained from the LandScan population dataset.

The maximum KDE values and average PRKD values between September 26, 2014 and November 5, 2014 for geographic boundaries of four representative cities (i.e. Dallas, New York City, Atlanta, and Los Angeles) are extracted from the KDE and PRKD output respectively. Dallas and New York City are selected because they are the places where Ebola cases were confirmed. Atlanta is a famous international transport hub, like New York City, which has very large mobile population. Los Angeles is now the second largest city in the US, which is smaller than New York City and larger than Dallas. Thus, Atlanta and Los Angeles can be used as reference cities when analyzing the maximum KDE and average PRKD values of Dallas and New York City.

3. Results

Explore temporal characteristics of Ebola-related tweets

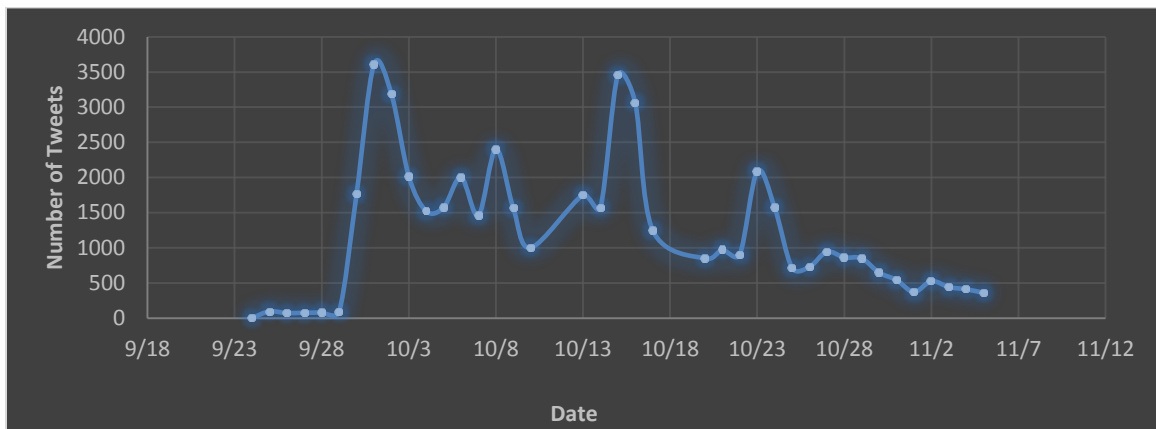


Figure 1: Changes in the number of Ebola-related tweets between Sept.25 and Nov. 5, 2014

A curve in Figure 1 shows the changes in the number of tweets regarding Ebola from September 25, 2014 to November 5, 2014. Before Ebola virus landed on the U.S., American public lacked concern about Ebola even though Ebola-epidemic-situation has been very serious in West Africa. However, with Ebola virus beginning to attack the U.S., Ebola information quickly "infected" American Twitter users. The number of Ebola-related tweets rocketed to 1764 on September 30, 2014 and reached 3607 on October 1, 2014. From October 2, 2014 to October 24, 2014, American Twitter users kept concerned about Ebola. It can be clearly seen from Figure 1 that four peaks of the curve exist on October 1, 2014, October 8, 2014, October 15, 2014, and October 23, 2014. October 1, 2014 is the next day after the first Ebola case was confirmed in the US. October 8, 2014 is the day when the first Ebola patient died. October 15, 2014 is the day when the third Ebola case was reported. October 23, 2014 is the day when the fourth Ebola case was confirmed. With the epidemic situation controlled and no more Ebola cases emerging, the number of Ebola-related tweets progressively decreased after October 25, 2014 and went down to 358 on November 5, 2014.

Explore spatial characteristics of Ebola-related tweets

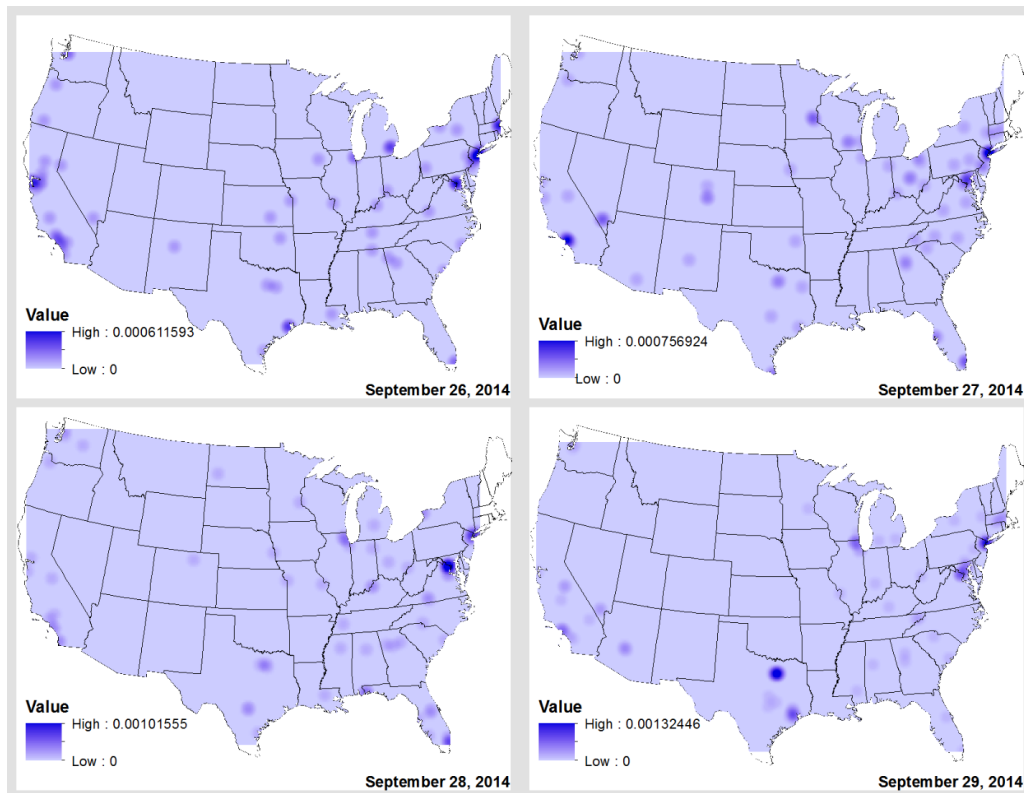


Figure 1: The KDE raster maps of Ebola-related tweets between Sept. 26, 2014 and Sept. 29, 2014 in the contiguous U.S.

The KDE maps (Figure 2) exhibit probability densities of posting Ebola-related tweets across the US. From September 26, 2014 to September 29, 2014, the kernel density in most areas of the US is very low (mostly smaller than 0.0016). Relatively larger KDE

values only emerge in a few metropolises, such as New York, Washington D.C., Los Angeles, Chicago, and Dallas (Figure 2). Yet, on September 30, 2014 with the first Ebola case confirmed in Dallas the maximum KDE values remarkably increase. The highest KDE value (0.006288) appeared in Dallas that is 95.8% larger than the second highest KDE value (0.003212) emerging in New York City (Figure 3). With the news that Ebola virus has emerged in the US spread, on October 1, 2014, the KDE values kept increasing across the whole country and the largest KDE values of that day (i.e., 0.007881) still existed in Dallas. After October 1, 2014, the KDE values across the whole country stopped to increase until the first Ebola patient died on October 8, 2014. On October 8, 2014, the maximum KDE value in Dallas where the patient died even reached 0.0261, which is 25.62% higher than the maximum KDE in New York City. From October 8, 2014 to October 15, 2014 when the third Ebola case was confirmed, the largest KDE value always emerged in Dallas. Yet, after October 15, 2014 KDE values in metropolitan area of New York increased dramatically, and after October 20, 2014 the largest KDE value emerged in New York City but not in Dallas. On October 23, 2014 when the fourth Ebola case was confirmed, the largest KDE value in New York City reached 0.0639, which is 93.60%, 95.65%, and 90.57% higher than the maximum KDE values in Dallas, Atlanta, and Los Angeles respectively (Figure 3).

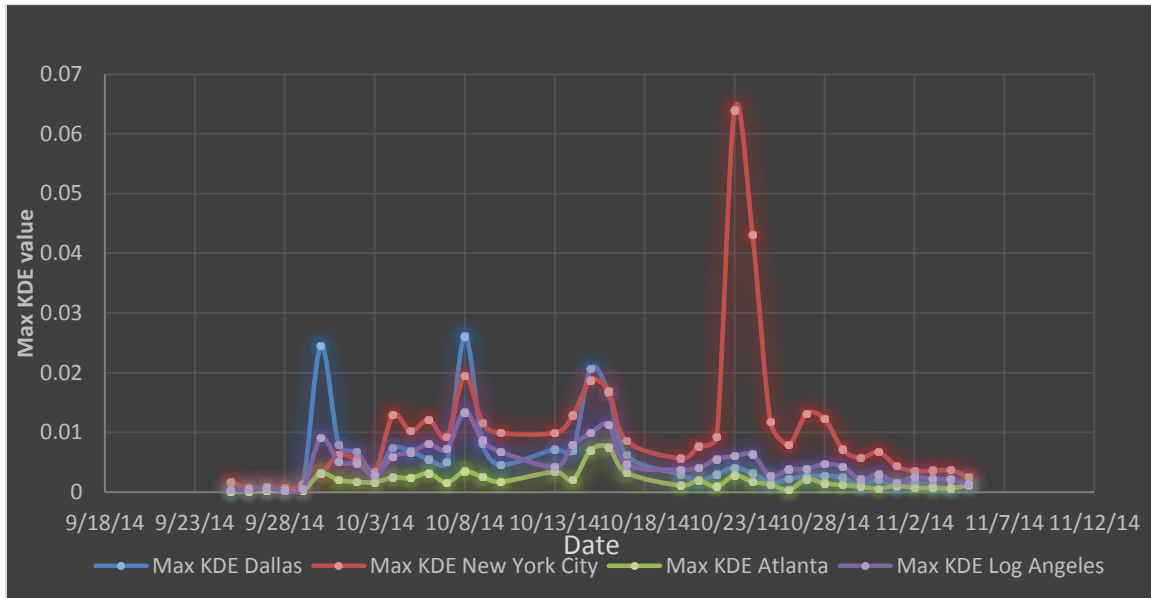


Figure 2: The maximum KDE values of Ebola related tweets for Dallas, New York City, Atlanta, and Los Angeles between September 25, 2014 and November 5, 2014.

However, Dallas and New York are both cities with very large population. Even though Ebola-related events did not occur in the two cities, very large KDE values are still likely to emerge in the cities. Consequently, we further analyzed PRKD for Ebola-related tweets. It can be found that spatio-temporal changing patterns of PRKD are almost the same with those of the KDE. The largest average PRKD values also emerged in the regions where Ebola-related events occurred (Figure 4). An apparent difference from the KDE maps is that in urban core regions very large population densities lead to

PRKD values smaller than those in urban and suburban regions (Figure 5 and Figure 6). However, this phenomenon of relatively small PRKD values in urban core regions does not impede us to identify places where Ebola-related events occurred. In Figure 5(a-c), it is clear to identify that Ebola-related events occurred in Dallas while in Figure 5(d-f), Ebola-related event can also be clearly detected occurring in Dallas because of their notably large PRKD values. Similarly, in Figure 6d and Figure 6h the maximum KDE value and the largest average PRKD value both emerge in New York City.

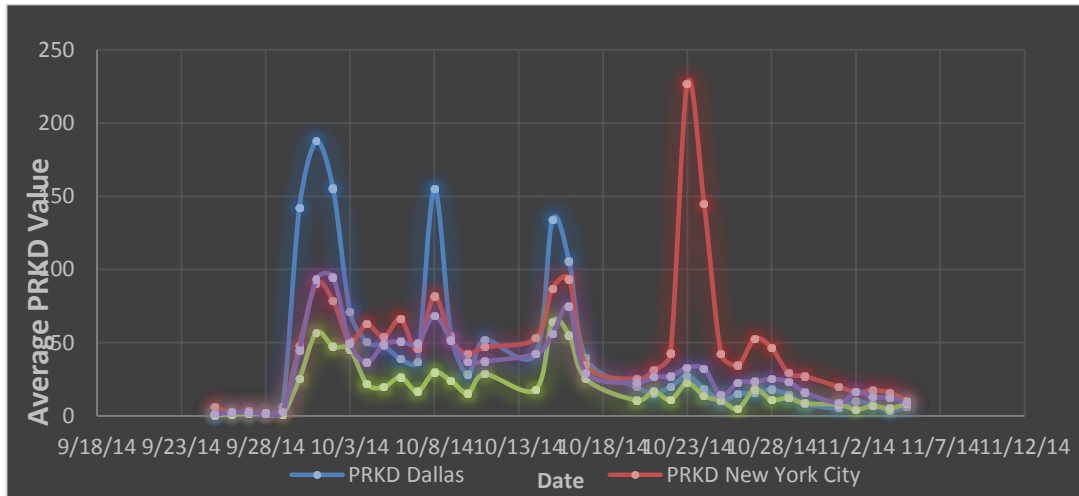


Figure 3: The average PRKD values of Ebola related tweets for Dallas, New York City, Atlanta, and Los Angeles between September 25, 2014 and November 5, 2014.

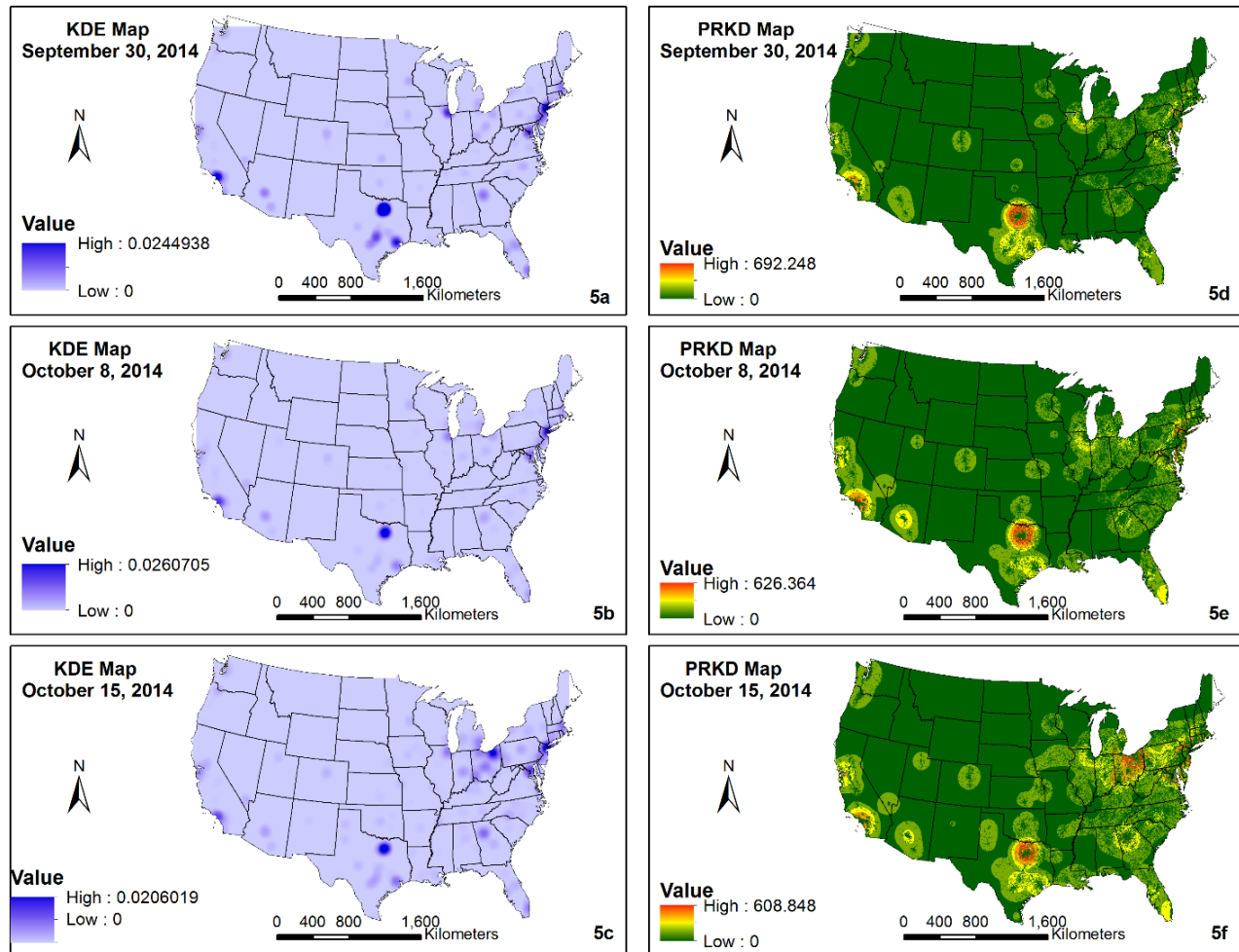


Figure 4: The KDE and PRKD maps of Ebola related tweets on September 30, 2014, October 8, 2014, and October 15, 2014.

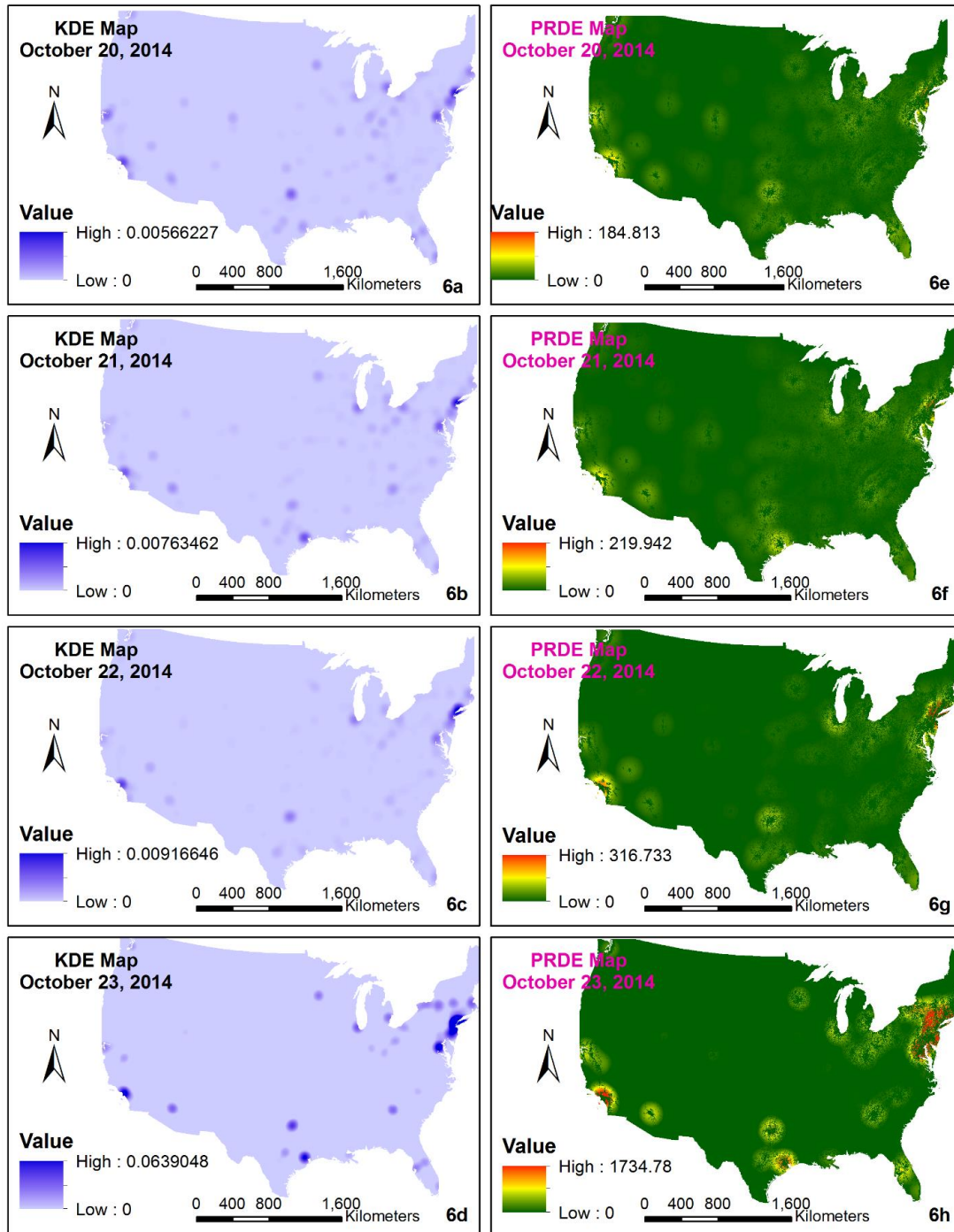


Figure 6: The KDE and PRKD maps of Ebola related tweets between October 20, 2014 and October 23, 2014.

4. References

- Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. 2013. Faking Sandy: characterizing and identifying fake images on twitter during hurricane sandy. In Proceedings of the 22nd international conference on World Wide Web companion, pp. 729-736.
- Kent, J.D., & Capello, H.T. 2013. Spatial patterns and demographic indicators of effective social media content during the Horseshoe Canyon fire of 2012. *Cartography and Geographic Information Science*, 40(2), 78-89.
- Parzen, E. 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3), 1065-1076.
- Sakaki, T., Okazaki, M., & Matsuo, Y. 2010. Earthquake shakes Twitter user: real-time event detection by social sensors. In Proceedings of the 19th international conference on World Wide Web, pp. 851-860.
- Signorini, A., Segre, A.M., & Polgreen, P.M. 2011. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS one*, 6(5), e19467.