Understanding the Prevalence of Drinking Water Service Disruption through Large-Scale Analysis of News Articles and Social Media

- Christine Kirchhoff, College of Engineering
- Shomir Wilson, College of Information Sciences and Technology

Abstract: Drinking water systems are vulnerable to disruption from extreme events including hurricanes, heavy rain and wind, drought, wildfire, and extreme cold (Nakamura and Brooks 2021; Proctor et al. 2020; Waldrop 2022), and from poor management, delayed maintenance, and age. These disruptions often result in drinking water advisories which require residents to boil their tap water before use, to not consume their tap water, or to not use their tap water (Public Notice Rule 141.202). Disruptions in essential drinking water services can cause ripple effects, with considerable social (Schulman et al. 2004, Pescaroli and Alexander 2015), health (Wade et al. 20, Quist et al. 2022) and economic consequences (Alzahrani and Collins 2022; Reich et al. 2017). Past research on drinking water advisories has predominantly focused on the quality of the notice (Wardlaw and Bruvold 1989; Harding and Anadu 2000; O'Shay et al. 2020), the cause (Hrudey et al. 2003; Hrudey et al. 2006; Kargiannis et al. 2009; Rimhanen-Finne et al., 2010; Robertson et al. 2009; Wallis et al. 2001), and on whether the public is aware and adheres to the recommendations (i.e., to boil water)(Jones-Bitton et al. 2016; Hyllestad et al. 2019; Kargiannis et al. 2009; O'Donnell et al. 2000; Ram et al. 2007; Rundblad et al. 2010; Vedachalam et al. 2016).

To date there has been no investigation that we know of to address the research question, What is the scope and scale of drinking water advisories in the United States (U.S.)? Moreover, anecdotal evidence suggests that drinking water disruptions may disproportionately affect poor and historically marginalized communities (Sneath 2021; Waldrop 2022) but this has not been investigated systematically. We seek to begin to fill this gap using public notices of drinking water advisories in print and broadcast media across the U.S. to characterize the frequency, cause including from extreme events, geographic location, water systems, and characteristics of populations affected by drinking water advisories over time. We will do this by developing a first-of-its-kind dataset of drinking water advisories across the U.S. The computational steps of this research will be (1) using web search and web crawling methods to gather textual artifacts (e.g., announcements, news articles) about drinking water advisories from official sources; (2) using semi-supervised machine learning methods to bin the articles into sets that represent discrete events; and (3) using information extraction and text classification techniques to label each artifact with key metadata, including dates and type of advisories, locations, responsible authority, and causes. This pilot research will lead to an NSF proposal for a more comprehensive collection and analysis of drinking water service disruptions.