Evidence-Based Health and Environmental Policies and the Potential Mismatch with Citizens' Perceptions: A Data Science Perspective

Nathalie de Marcellis-Warin, PhD
Full Professor, Polytechnique Montreal
CEO and Fellow, CIRANO Research Center
Visiting scientist, Harvard T.H. Chan School of Public Health
Nathalie.demarcellis-warin@polymtl.ca

Ann Backus, MS
Director of Outreach, Department of Environmental Health, Harvard T.H. Chan School of Public Health

ASSA 2019

Project funded in part by the HSPH-NIEHS Center for Environmental Health which is funded by The National Institute of Environmental Health Sciences P30 ES 000002
INTRODUCTION: DATA SCIENCE AND EVIDENCE-BASED PUBLIC POLICY

- Integrating large quantities of data from multiple, disparate sources can create new opportunities to understand complex questions.

- Currently, efforts are under way to develop methods to reliably integrate data from sources:
  - that are **not traditionally used** such as electronic medical files, data used in peer-reviewed publications and crowd-based sources
  - with **location information** such as geospatial datasets and geolocalized tweets.
Visualizations of geospatial datasets and crowd-based sources (geolocalized tweets, etc) could help to inform a specific decision and help communicate the results of an analysis.
The purpose of 1) data integration and 2) data visualization for environmental health research and decision-making is to improve public health by monitoring environmental exposures and health outcomes.

CASE STUDY: [FrackMap Project] Natural Gas and Shale Gas Extraction in the US
A data science perspective

- Natural gas and shale gas extraction operations create several social, environmental and economic challenges for local communities.
- Energy-based companies highlight the economic opportunity of such operations (local, regional, state, and national level) and scientific studies point out a vast array of potential and proven risks: ecological, seismic, public health, occupational health, etc.

1) Integrating data
2) Visualizing data
Integrating large quantities of data from multiple, disparate sources can create new opportunities to understand complex environmental health questions.

- Natural gas and shale gas extraction operations data
- Peer-reviewed publications on potential and real impacts of hydraulic fracturing on health and environment
- Tweets about #shalegas and #fracking (public perception)
CASE STUDY: Natural Gas and Shale Gas Extraction in the US
A data science perspective

1) Integrating data

Natural gas and shale gas extraction operations data*:
- shale formations’ locations
- wells’ locations
- horizontal legs’ locations
- production by state
- permits by state and by wells
- reports of specific chemical used by wells
- regulations by state, etc.

* Publicly available data
U.S. natural gas production hit a new record high in 2018

U.S. annual natural gas production (1940-2018)

billion cubic feet per day

120

100

80

60

40

20

0


gross withdrawals
marketed production
dry production

Source: U.S. Energy Information Administration, Monthly Crude Oil, Lease Condensate, and Natural Gas Production Report, Natural Gas Monthly

https://www.eia.gov/todayinenergy/detail.php?id=42337
Note: Fracking was disallowed in three states – New York, Vermont, and Maryland – due to the risk of contaminated drinking water (Boersma & Johnson, 2012).
CASE STUDY: Natural Gas and Shale Gas Extraction in the US
A data science perspective

1) Integrating data

More than 1000 peer-reviewed publications with datasets* on potential impacts about:

• water quality,
• air quality,
• induced seisms,
• publich health,
• etc

* with specific information about the location
METHODOLOGY (peer-reviewed articles database)

- Peer-reviewed publications from January 2005 to November 2019 about environmental and health impacts in the US
- Systematically searched and screened
- Databases: PubMed, MEDLINE, ScienceDirect, Scopus, Web of Science, Proquest, Google Scholar, etc
- Key research terms: water impacts (water usage, wastewater, water quality (ground water), Air pollution, Climate change (greenhouse gases, large scale impacts), Ecological impacts (forestry, fauna and flora), Health (public health and occupational exposure), Seismicity, etc.
- Data location (in the title, abstract, keywords): State, County, City, Shale Play, River, Lake,…

1000-ish peer-reviewed publications with geographical data
Microearthquakes (that is, those with magnitudes below 2) are routinely produced as part of the hydraulic fracturing (or “fracking”) process used to stimulate the production of oil, but the process as currently practiced appears to pose a low risk of inducing destructive earthquakes. More than 100,000 wells have been subjected to fracking in recent years, and the largest induced earthquake was magnitude 3.6, which is too small to pose a serious risk. Yet, wastewater disposal by injection into deep wells poses a higher risk, because this practice can induce larger earthquakes. For example, several of the largest earthquakes in the U.S. midcontinent in 2011 and 2012 may have been triggered by nearby disposal wells. The largest of these was a magnitude 5.6 event in central Oklahoma that destroyed 14 homes and injured two people. The mechanism responsible for inducing these events appears to be the well-understood process of weakening a preexisting fault by elevating the fluid pressure. However, only a small fraction of the more than 30,000 wastewater disposal wells appears to be problematic—typically those that dispose of very large volumes of water and/or communicate pressure perturbations directly into basement faults.
Peer-reviewed articles database
CASE STUDY:
Natural Gas and Shale Gas Extraction in the US
A data science perspective

1) Integrating data

- More than 65,000 geolocalized tweets about #shalegas and #fracking (public perception)
Twitter and Fracking

• The domain of public opinion, political agenda, and the controversy of fracking is nowadays a well-studied phenomenon, where public attitudes were massively influenced online in social medias in addition to the traditional news media (Hopke & Simis, 2015).

• The hashtag #fracking can be used to capture the viral messages related to anti-fracking sentiments sent by prominent actors or opinion leaders. The support groups for fracking use other hashtags such as natural gas (#natural-gas) or shale oil (#shale-oil) (Sharag-Eldin, Ye, & Spitzberg, 2018).

• Social medias such as Twitter allow new forms of activism, for instance the organization and promotion of an environmental movement centered on a transnational day of action calling for a ban on hydraulic fracturing: the Global Frackdown (Hopke, 2015).
METHODOLOGY (geolocalized tweets dataset)

1 Tweet = 140 characters maximum, including keywords-hashtags (#) (+ image or video or text, etc ) (from 2018 = 280 characters)

A tweet contains more than **40 elements in its metadata:**
- the name of the user that sent the message,
- its geolocation (if activated),
- the time the message was sent,
- the content of the message,
- how many times the message has been liked, etc.

Moreover, metrics such as the sentiment associated with a message or how many times it has been retweeted provide additional information.
METHODOLOGY (geolocalized tweets dataset)

Using hashtags of the keywords used for Biblio

- #Fracking #FrackingWasteWater #FrackQuake #EarthQuake
- #ShaleOil #ShaleGas
- #MarcellusShale #UticaShale #BarnettShale #BakkenShale #EagleFordShale

Data from Harvard Center Geographic Analysis: « One Billion Tweets Project » (2012-2015)

- Harvard CGA Geolocated Archive / Geotagged Tweets
- List of # and keywords

Use of the Nuance-R technological Platform (PI: Warin, T. 2015)
We access the Twitter REST API with the streamR R package [Barbera, 2018] and selected #

➢ 65,000 tweets
METHODOLOGY (geolocalized tweets dataset)
Content and Sentiment Analysis:

The first step of the analytical analysis is to tidy our dataset following Hadley Wickham’s description [Wickham, 2014]: "each variable is a column, each observation is a row, each type of observational unit is a table".

In order to associate a sentiment score to each tweet we manipulate our dataset in order to remove all links from the messages, then tokenize each message and finally we remove all stopwords following Silge and Robinson (2017) approach.

We compare the results of the sentiment analysis of the messages associated to each # with the lexicon [Hu et Liu, 2004].
Opinion Formation: Most used # hashtags

#Fracking #Shale
#Utica
#Jobs # Job #Kellyjobs #Hiring #Career
#Pipeline #Oil #Energy
#Physician #Nursing #Transportation
Opinion Formation: "Relevant" words by count

2012 to 2015

2015
Sentiment Analysis: Classification by Polarity

![Sentiment Analysis of Tweets](chart.png)
CASE STUDY: Natural Gas and Shale Gas Extraction in the US A data science perspective

2) Visualising data

Visual communication of data from multiple, disparate sources can create new opportunities to understand complex environmental health questions.

- Natural gas and shale gas extraction operations data: wells locations, etc.

- More than 1000 peer-reviewed publications on potential and real impacts of hydraulic fracturing on health and environment with data locations

- More than 65 000 geolocalized tweets about #shalegas and #fracking (public perception)
The FrackMap brings together a range of fracking related datasets

Data: oil and gas permits, shale formations, horizontal legs, etc.

Harvard WorldMap a public domain collaborative mapping platform
http://worldmap.harvard.edu/maps/FrackMap
Algorithms to gather geographical data within articles

Article mentions state

Does the article mention shale?

Yes  No

Get a map of the state with shale plays information. Does the state only have one shale within its borders?

Yes  No

Pinpoint the center of the Shale. Association of that shale to all such state missing a shale in the database.

E.g.: Fayetteville (AR); Monterey (CA); Antrim (MI).

Pinpoint the center of the portion of the shale within that state.

E.g.: Bakken (MT, ND); Niobrara (CO, NE, WY); Marcellus (NY, OH, PA, WV).

Get a map of the state with shale plays information. Does the state only have one shale within its borders?

Yes  No

Pinpoint the center of the shale. Association of that shale to all such state missing a shale in the database.

E.g.: Fayetteville (AR); Monterey (CA); Antrim (MI).

Pinpoint the state capital.

E.g.: Texas (Barnett, Eagle Ford, Haynesville, Permian) is pinpointed on Austin. Colorado (Wattenberg, Piceance, Niobrara) on Denver.
Map, by state, peer-reviewed literature about potential environmental and health issues and impacts associated with U.S. shale gas plays
Mapping the tweets, 2012
Mapping the tweets, 2015
Mapping the tweets, water impacts
Mapping the tweets, earthquake + frackquake
Conclusion

Integrating and visualizing large quantities of data from multiple sources can create new opportunities to understand complex questions and could help communicate.

FrackProject is an innovative tool to integrate data and communicate through maps and interactive data visualization.

FrackProject could help regulators and industry to implement best risk management practices and invent safer practices.

- Twitter is an interesting platform:
  - to study opinion formation and the nature and pace of the spread of an information through Twitter conversations
  - The conversation is more about #jobs, #jobs, #jobs... which contradicts the evidence.