Supporting Information

"Exposure assessment using secondary data sources in unconventional natural gas development and health studies"

Kirsten Koehler, PhD[1]; J. Hugh Ellis, PhD[1]; Joan A. Casey[2]; David Manthos, BA[3]; Karen Bandeen-Roche, PhD[4]; Rutherford Platt, PhD[5]; and Brian S. Schwartz*, MD, MS[1,6,7]

[1]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA; [2]SkyTruth, Shepherdstown, WV, USA; [2]Division of Environmental Health Sciences, University of California at Berkeley School of Public Health, Berkeley, California, USA; [3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA; [4]Department of Environmental Studies, Gettysburg College, Gettysburg, Pennsylvania, USA; [5]Department of Epidemiology and Health Services Research, Geisinger Health System, Danville, Pennsylvania, USA; [6]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA

Corresponding author(*): Brian S. Schwartz, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Room W7041, Baltimore, MD, 21205. E-mail: bschwar1@jhu.edu.

Number of pages: 5

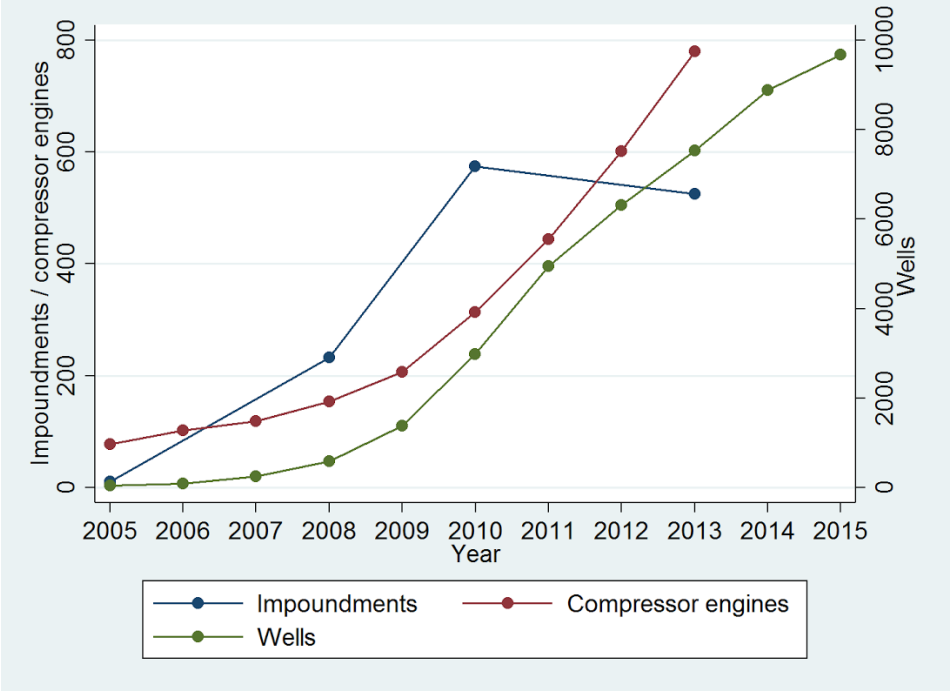
Number of figures: 1

Number of tables: 1

1. **Impoundment and flaring data access**

   The impoundment data can be accessed on SkyTruth's website at http://skytruth.org/2014/10/pa-drilling-impoundments-2005-2013/, and the flaring data can be accessed at http://skytruth.org/viirs/.

2. **Figure S1.** Total number of drilled unconventional natural gas wells and operating unconventional natural gas related impoundments and compressor engines in Pennsylvania by year.



3. Information regarding impoundments

Information on impoundment location and sizes was obtained in partnership with SkyTruth, which created a collaborative image analysis application on their website (skytruth.org) that displayed aerial imagery collected by the USDA National Agricultural Imagery Program[1] of the one square kilometer area around UNG wells from the summers of 2005, 2008, 2010 and 2013 (Abstract Art). Trained volunteers and staff looked at aerial imagery and identified and outlined impoundments. Volunteers contributing to the project were trained using an interactive tutorial, once they achieved an accuracy of 100% on the training data they were able to progress on and to contribute to project. Each image was reviewed by no less than three staff or ten volunteers, 80% agreement was required for volunteers and 66% agreement was used internally, and assignments were validated by a GIS analyst before inclusion in the final dataset.[2]

To estimate an installation and removal date for each impoundment, we used a trend analysis of Landsat data to identify sudden spectral changes in the grid cell that contained each impoundment.[3] To do so, we compiled all available Landsat 5, 7, and 8 surface reflectance imagery with < 30% cloud cover for the years 2000-2015, a total of 754 images across four Landsat path/rows. For each impoundment location, we masked remaining clouds and then interpolated a monthly time series for the near infrared band and the normalized difference vegetation index (NDVI). We used the Breaks for Additive Season and Trend package in *R* to identify discrete breaks in the time series after the removal of seasonal effects.[4] Based on the direction, magnitude, and timing of the time series breaks, we identified approximate dates of creation and removal of impoundments. We verified estimates for a sample of impoundments by comparing Landsat-derived dates to photointerpretation-derived dates using historical imagery on Google Earth.

Temporality of the satellites introduced some measurement error into our impoundment estimates. Landsat 5 was retired in 2012 and Landsat 8 images the entire planet every 16 days (in a 7-day offset from Landsat 7, see https://landsat.usgs.gov/landsat-8). Furthermore, cloud cover and Landsat 7 SLC-off gaps introduced additional error. Therefore, our estimates of impoundment installation and removal dates could have differed from the truth by more than a month in some cases. We have no reason to expect this error is differential with respect to exposure status over time, so the error likely biased our results towards the null.

4. Information regarding flaring events

4.1 Methods to identify flaring

We also identified flaring events using detections recorded at night by the Visible Infrared Imaging Radiometer Suite on the Suomi NPP satellite operated by the National Oceanic and Atmospheric Administration (NOAA). We identified detections in Pennsylvania with a temperature >1773ºK and <5273.15ºK (excluding temperatures of 1810ºK, which NOAA used to identify detections where it is not possible to estimate the temperature) from September 9, 2012 to August 3, 2015. However, the PCA did not identify these events as important and they were not included in the final UNGD metric. We did not incorporate flaring events into this analysis because we did not have information on flaring events before 2013 and only four locations had flaring events identified in 2013.

3.2 Results related to flaring data

Between September 2012 and August 2015, we identified 1,174 flaring observations on 380 days. After grouping flares within 150m, we identified flares at 216 locations (**Figure 1**). At 114 locations (53%), the flaring event was observed on one day, and at the remaining 102 sites, there was a median of 115 days from the first to last flaring event. In the Marcellus shale, flares tend to run continuously for weeks during peak periods of activity at the well pad.

The flaring data had two related limitations. First, we likely underestimated the number of flaring events because we could not identify flaring events on cloudy nights. Second, the Suomi NPP satellite passes over Pennsylvania twice every 24 hours, once during daylight and once at night (see https://www.nasa.gov/mission_pages/NPP/main/index.html and https://www.skytruth.org/viirs/). We only used the nighttime imagery to identify flares. Therefore, if flaring events lasted only for a short time period (<15 hours or so, during daylight hours), we might not have captured them in our dataset. However, the high-temperature of the flares and extended, continuous duration of flaring effectively excludes the possibility of a false positive.[5] The greatest uncertainty would be the full duration of the flare, with several days of possible missed detections at either end of the flaring event. More research is necessary to effectively determine when a null detection was due to atmospheric conditions or a true negative. This limitation could mean that we captured only a subset of flares that occurred in Pennsylvania during the study period.

Table S1. Results of PCA with Percentage of Variation Explained by Component 2 and Component 2 Loadings

Date	Proportion of variance explained by component 2	Component 2 loadings					
			Well metrics				
		Compressor engine metric	Pad	Drilling	Stimulation	Production	Impoundment metric
1/1/2005	0.20	0.83	a	a	a	-0.19	0.59
7/1/2005	0.15	-0.15	-0.46	0.33	a	-0.08	0.07
1/1/2006	0.07	0.81	-0.23	a	a	0.58	b
7/1/2006	0.04	0.86	-0.24	0.06	0.46	0.44	b
1/1/2007	0.11	0.08	0.83	-0.45	-0.32	0.01	b
7/1/2007	0.18	-0.13	0.97	-0.18	0.05	-0.13	b
1/1/2008	0.12	-0.09	0.21	-0.11	-0.34	-0.23	0.88
7/1/2008	0.27	-0.31	0.49	0.53	-0.43	-0.30	0.33
1/1/2009	0.29	-0.35	-0.12	0.54	-0.41	0.63	b
7/1/2009	0.19	0.90	0.02	-0.14	0.05	-0.40	b
1/1/2010	0.22	0.56	0.49	-0.32	-0.46	-0.29	0.20
7/1/2010	0.11	0.78	-0.17	-0.05	-0.24	-0.46	0.29
1/1/2011	0.10	0.87	-0.05	-0.33	0.01	-0.36	b
7/1/2011	0.08	0.84	-0.36	-0.23	0.13	-0.29	b
1/1/2012	0.07	0.86	-0.33	-0.24	0.08	-0.29	b
7/1/2012	0.11	0.76	-0.63	0.05	0.02	-0.15	b
1/1/2013	0.10	0.58	-0.39	-0.34	-0.24	-0.10	0.57
7/1/2013	0.09	-0.48	0.37	0.08	0.52	0.12	-0.59

[a] All grid points had a value of zero for this variable on this date, and variables with zero variance were dropped from PCA.
[b] Impoundment data was only available in 2005, 2008, 2010, and 2013

5. References

1. United States Department of Agriculture (USDA). National Agriculture Imagery. https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/index. Accessed 26 Dec 2016.
2. Wurster K. SkyTruth FrackFinder PA 2005-2013 Methodology. 2014; https://github.com/SkyTruth/CrowdProjects/tree/master/Data/FrackFinder/PA. Accessed 21 Jan 2018.
3. Platt RV, Manthos D, Amos J. Estimating the creation and removal date of fracking ponds using trend analysis of Landsat imagery. *Environ Manag.* 2018:1-11.
4. Verbesselt J, Hyndman R, Newnham G, Culvenor D. Detecting trend and seasonal changes in satellite image time series. *Remote Sens Environ.* 2010;114(1):106-115.
5. Saberi P, Propert KJ, Powers M, Emmett E, Green-McKenzie J. Field survey of health perception and complaints of Pennsylvania residents in the Marcellus Shale region. *Int J Environ Res Public Health.* 2014;11(6):6517-6527.