Journal of Contaminant Hydrology 195 (2016) 23–30

Contents lists available at ScienceDirect

Journal of Contaminant Hydrology

journal homepage: www.elsevier.com/locate/jconhyd

# Searching for anomalous methane in shallow groundwater near shale gas wells

Zhenhui Li [a,*], Cheng You [b], Matthew Gonzales [c], Anna K. Wendt [c], Fei Wu [a], Susan L. Brantley [c,*]

[a] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, United States
[b] Department of Statistics, The Pennsylvania State University, University Park, PA 16802, United States
[c] Earth and Environmental Systems Institute and Department of Geosciences, The Pennsylvania State University, University Park, PA 16802, United States

ARTICLE INFO

*Article history:*
Received 22 June 2016
Received in revised form 22 October 2016
Accepted 26 October 2016
Available online 28 October 2016

*Keywords:*
Hydraulic fracturing
Methane
Groundwater
Data mining

ABSTRACT

Since the 1800s, natural gas has been extracted from wells drilled into conventional reservoirs. Today, gas is also extracted from shale using high-volume hydraulic fracturing (HVHF). These wells sometimes leak methane and must be re-sealed with cement. Some researchers argue that methane concentrations, *C*, increase in groundwater near shale-gas wells and that "fracked" wells leak more than conventional wells. We developed techniques to mine datasets of groundwater chemistry in Pennsylvania townships where contamination had been reported. Values of *C* measured in shallow private water wells were discovered to increase with proximity to faults and to conventional, but not shale-gas, wells in the entire area. However, in small subareas, *C* increased with proximity to some shale-gas wells. Data mining was used to map a few hotspots where *C* significantly correlates with distance to faults and gas wells. Near the hotspots, 3 out of 132 shale-gas wells (~2%) and 4 out of 15 conventional wells (27%) intersect faults at depths where they are reported to be uncased or uncemented. These results demonstrate that even though these data techniques do not establish causation, they can elucidate the controls on natural methane emission along faults and may have implications for gas well construction.



## 1. Introduction

In the U.S.A., the usage of natural gas has increased markedly due to new techniques in developing gas directly from shale. Since 2014, this so-called "unconventional" resource has been estimated to comprise about 50% of the total proven U.S.A. gas reserves (U.S. Energy Information Administration, 2014). Extraction of gas from shale has become possible due to improvements in directional drilling and high-volume hydraulic fracturing (HVHF) (Vidic et al., 2013a). The rapid development in the use of HVHF in the U.S.A. since the 1990s has sometimes led to water quality impacts that have caused concern, including leakage of methane out of gas wells due to well integrity problems (Vidic et al., 2013a). Such problems have been particularly controversial in the Marcellus gas play because this shale formation underlies 8 highly populous northeastern states. One state regulator, the Pennsylvania Department of Environmental Protection (PA DEP), reported, for example, that the most common type of water quality impact related to oil/gas activity by companies developing "unconventional" wells – i.e. wells completed with HVHF – is methane contamination (Brantley et al., 2014). The frequency of well integrity problems (Brantley, 2014) for wells completed with or without HVHF – i.e., "fracking" – is important given that leakage into drinking water resources entails explosion hazards when concentrations approach 10 ppm and methane in groundwater can result in secondary contamination (Vidic et al., 2013a). In addition, eventual release of methane into the atmosphere increases greenhouse warming (Howarth et al., 2011). According to PA DEP records, 3.4% of gas wells were cited for well construction problems before 2013 (Vidic et al., 2013a). Of these, 16 wells (0.24%) were cited for allowing gas to migrate into groundwater. This methane leakage rate in the Marcellus play may have changed with time as operators learned better practices (Brantley, 2014). However, the leakage rate, which is difficult to quantify, has become controversial for HVHF because some have claimed that natural gas leaks more readily from wells in unconventional formations than from "conventional" wells (Howarth et al., 2011; Ingraffea et al., 2014).

One way to investigate leakage is to determine if the concentration of methane in groundwater, *C*, varies with distance from gas wells. However, such studies depend on how many water wells are investigated. For example, an early investigation concluded that *C* increased in ~60 waters sampled within 1 km of unconventional wells located in Pennsylvania (U.S.A.) (Osborn et al., 2011). This claim has been disputed (Davies, 2011; Molofsky et al., 2011; Schon, 2011; Jackson et al., 2011; Molofsky et al., 2013) at least partly because *C* can be high in groundwater due to natural processes (Reese et al., 2014; Baldassare et al., 2014). In a second investigation of the same area with 141 samples, *C* once again was observed to increase near gas wells (Jackson et al., 2013a). However, both the original and extended studies included observations

* Corresponding authors.
*E-mail addresses:* jessieli@ist.psu.edu (Z. Li), sxb7@psu.edu (S.L. Brantley).

http://dx.doi.org/10.1016/j.jconhyd.2016.10.005
0169-7722/

around Dimock PA where investigators have concluded a few gas wells contaminated 18 water supply wells in the early days of shale gas development (Brantley et al., 2014). In contrast, an analysis of >11,000 water samples in northeastern PA revealed no correlation between *C* and proximity to unconventional wells (Siegel et al., 2015). Apparently, higher *C* values may have been present in the larger dataset but not detectable because of the large number of non-impacted groundwater samples. None of these datasets have been released in entirety because of concerns about homeowner confidentiality.

In this paper we analyze a newly published data set from the PA DEP (1690 water samples from shallow, private water wells) from Bradford County, Pennsylvania (Shale Network, 2015) to learn how to interpret environmental datasets of different size. We hypothesized that large datasets, on average, might mask contamination that could be observed in smaller datasets. We also sought to understand the importance of conventional versus unconventional wells and the effect of local geology on methane emission. Our analysis focused on five townships where impacts to groundwater from methane were reported. We developed strategies to use large groundwater datasets to highlight and understand possible sites of methane emission with respect to local conditions. Our new technique relies on the use of large datasets and should be broadly applicable to other environmental data where patterns in distribution of contamination may allow for better environmental practices.

## 2. Analyzed data

To determine environmental patterns using data mining requires the availability of large numbers of analyses. Large datasets generally require that environmental data be pooled from many sources. The strategy of using large datasets and data mining is therefore predicated on the assumption that fundamental patterns can be gleaned from large datasets even though such sets may be characterized by variable data quality. We implicitly test that proposition here.

The water samples we analyze were collected by independent environmental consultants paid by gas companies before drilling and measured in commercial analytical laboratories that support extensive quality control and assurance measures (see Suppl. Information). The analyses are released to the state regulator to protect the gas company from future liability if water issues are reported. Given this end use, biasing samples or analyses toward lower methane concentrations (for example by allowing volatilization) is likely to be counter-productive.

Water samples were collected prior to treatment, filtration or water softening using U.S. Geological Survey protocol. Samples were collected and analyzed in accordance with Pennsylvania code § 78.52 which states, "(c) The survey shall be conducted by an independent certified laboratory. A person independent of the well owner or well operator, other than an employee of the certified laboratory, may collect the sample and document the condition of the water supply, if the certified laboratory affirms that the sampling and documentation is performed in accordance with the laboratory's approved sample collection, preservation and handling procedure and chain of custody."

Following a data sharing agreement between PA DEP and Pennsylvania State University, we analyzed data from Bradford County for five townships (Fig. 1). No attempt was made to analyze variation in *C* with time at each location because very few water wells were sampled more than once. The waters were sampled from water wells (average depth 54 m; ranging from 2 to 250 m) before drilling the new gas wells over a period of a few years. However, because the water sampling generally occurred near already-drilled gas wells, the data were investigated here with respect to gas wells that had already been drilled in conventional or unconventional formations. Each water analysis (i.e., sample site) was paired with the closest *previously drilled* unconventional well using data on the PA DEP Oil and Gas Reporting Website as of April 2015 (Murphy, 2012). Distances were determined for the closest already-drilled well (i.e., spud date prior to water sampling) within both the targeted and nearby townships. Of the original 1240 unconventional wells considered for the region, sample sites were paired with 132 unconventional gas wells spudded from June 2008 through July 2012. Likewise, of the 113 conventional gas wells in the overall region, samples were paired with 15 conventional wells: 13 spudded between 1932 and 1983 but now abandoned, and two spudded in 2009 and still active. The number of analyses and wells included in this dataset is intermediate between the previously discussed published datasets (Osborn et al., 2011; Jackson et al., 2013a; Siegel et al., 2015) and this allowed us to test how the size of the dataset affects conclusions about methane migration. In addition, the dataset reported here is the only one published with locations (Shale Network, 2015). More details are described in the Suppl. information (SI).

## 3. Methods

We analyzed the full dataset and then used increasingly finer spatial resolution by employing the following steps. First, we plotted *C* versus the distance to the nearest already-drilled unconventional or conventional well for the entire dataset. We quantified the correlation between *C* (i.e., dependent variable y) and distance (i.e., independent variable x). However, many statistical measures are not applicable because of the multiple reporting limits (i.e., detection limits) (Siegel et al., 2015; Helsel, 2011). For example, Pearson correlation and linear regression are not suitable; furthermore, Spearman correlation is only suitable for data with one reporting limit. Therefore, we used three measures that are appropriate for censored data with multiple reporting limits: Kendall rank correlation, Akritas-Theil-Sen (ATS) regression, and logistic regression (see SI for more details).

We next subdivided our study area into three subregions (A, B and C), which were selected to produce three clusters largely delineated by townships, each with at least 350 analyses (Fig. 1). The correlation statistics were then re-calculated for samples collected within each subregion.

To learn to analyze subregions of these environmental data randomly, we then developed a new sliding window approach inspired by the spatiotemporal exploratory model (Fink et al., 2010). We scanned the whole region using a "sliding window" of size 5 km × 5 km that was stepped over the map in 200 m increments. For each sliding window observation at each location separated by 200 m, we tested for Kendall rank correlation for the data in the window. The window was marked as +1 if the correlation is significantly positive and −1 if significantly negative (significance level of 5%). A spatially-normalized significance value was assigned to each location as defined by the sum of all windows covering the location divided by the total number of windows covering the location. The spatially-normalized significance values, plotted every 200 m, were then used to generate correlation maps showing regions of higher positive or negative correlations.

With the correlation maps, we explored the relationship between hot spots and the underlying geologic structure using maps of known faults in the area. The hot spots are the locations showing negative correlations between methane concentration and distance to well, i.e., higher dissolved methane concentrations closer to the well. For the wells located near hot spots we also investigated the well characteristics (e.g., casing and cementing). Finally, because all the unconventional wells had not been hydraulically fractured by the time of water sampling, we also repeated our methodology on the subset of wells that were completed by HVHF prior to water sampling.

## 4. Results and discussion

Fig. 2 shows scatter plots of *C* versus distance to the nearest already-drilled unconventional well before and after log transforming the data. These plots are visually misleading because a high percentage of samples cluster near the reporting limits of 1, 5 and 26 ppb (Siegel et al., 2015). A binned plot of the same data (Fig. 3) documents that such