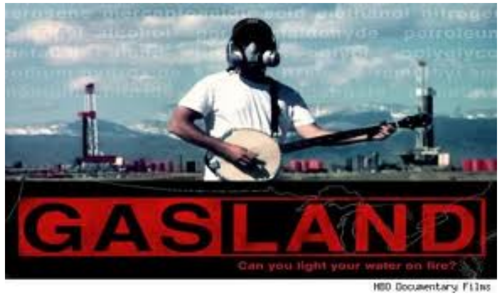
Statistical Analysis and Data Mining on Water Quality Data

Zhenhui (Jessie) Li[1], Cheng You[2], Fei Wu[1], Matt Gonzales[3], Sue Brantley[3]
[1]College of Information Sciences and Technology
[2]Department of Statistics
[3]College of Earth and Mineral Science

Public Concern of Water Contamination



The most often cited problem by the PA DEP (Dept of Environmental Protection) between 2008 and 2012 with respect to oil / gas impacts on drinking water systems is natural gas getting into ground water. Some of these incidents were reported in the media, leading to some public pushback against the use of hydraulic fracturing.

DEP's Investigation about Complaints of Water Contamination by Gas Wells

- From 2008 to 2012, PA DEP received ~1000 complaints about contamination in water wells
- 17% of incidents (161) were deemed caused by oil/gas, roughly split between conventional & unconventional
- From 2008 to 2012, 90 entities (households, churches, etc.) complained about drinking water irregularities and were told by the DEP that their drinking water was contaminated with natural gas. For these sites, a positive determination was made
  - There are 7000 gas wells in PA
  - For 90 sites, the oil and gas company could not refute that their activity had caused the contamination

Brantley et al., 2014, International Journal of Coal Geology 126 (2014) 140-156

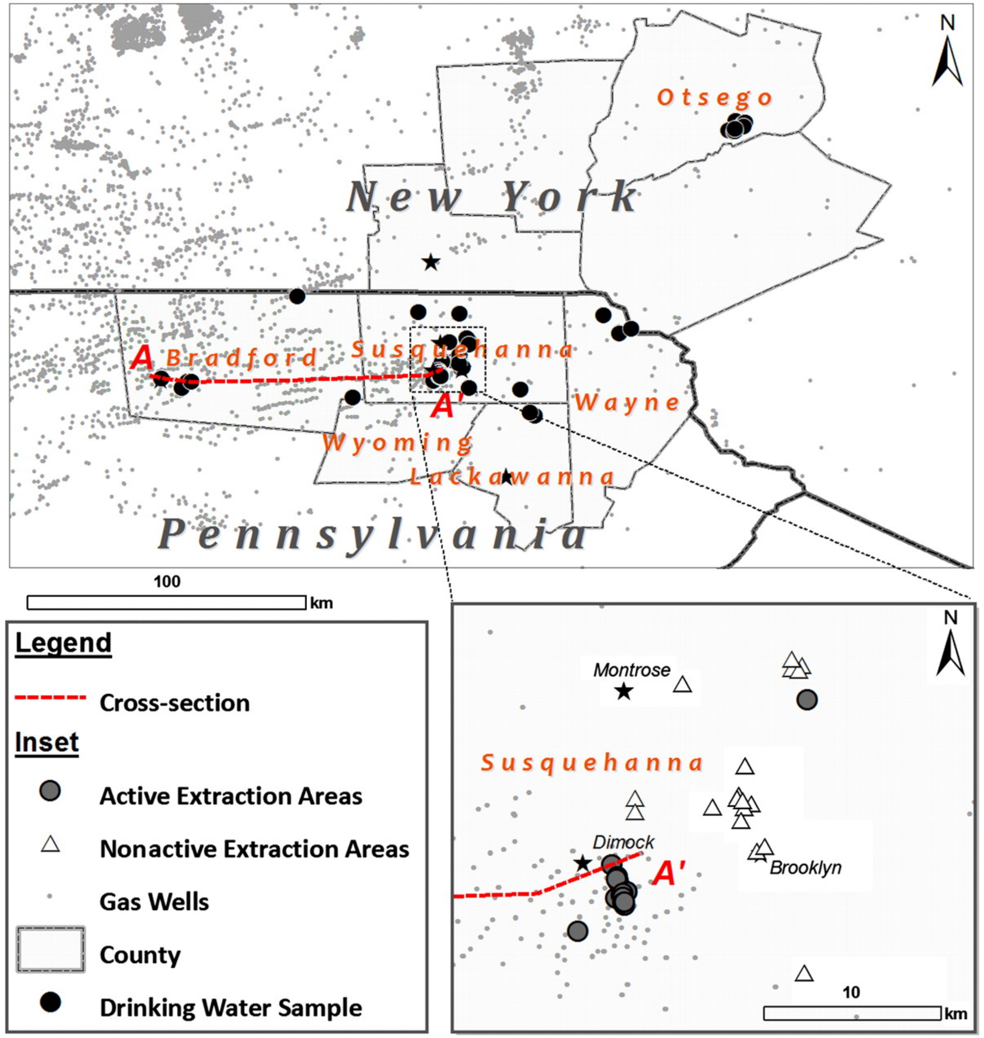
Academic Investigation about Methane

- "Our results show evidence for methane contamination of shallow drinking-water systems in at least three areas of the region and suggest **important environmental risks** accompanying shale-gas exploration worldwide." (Osborn et al., 2011)
- "In contrast to prior findings, we found **no statistically significant relationship** between dissolved methane concentrations in groundwater from domestic water wells and proximity to pre-existing oil or gas wells." (Siegel et al., 2015)

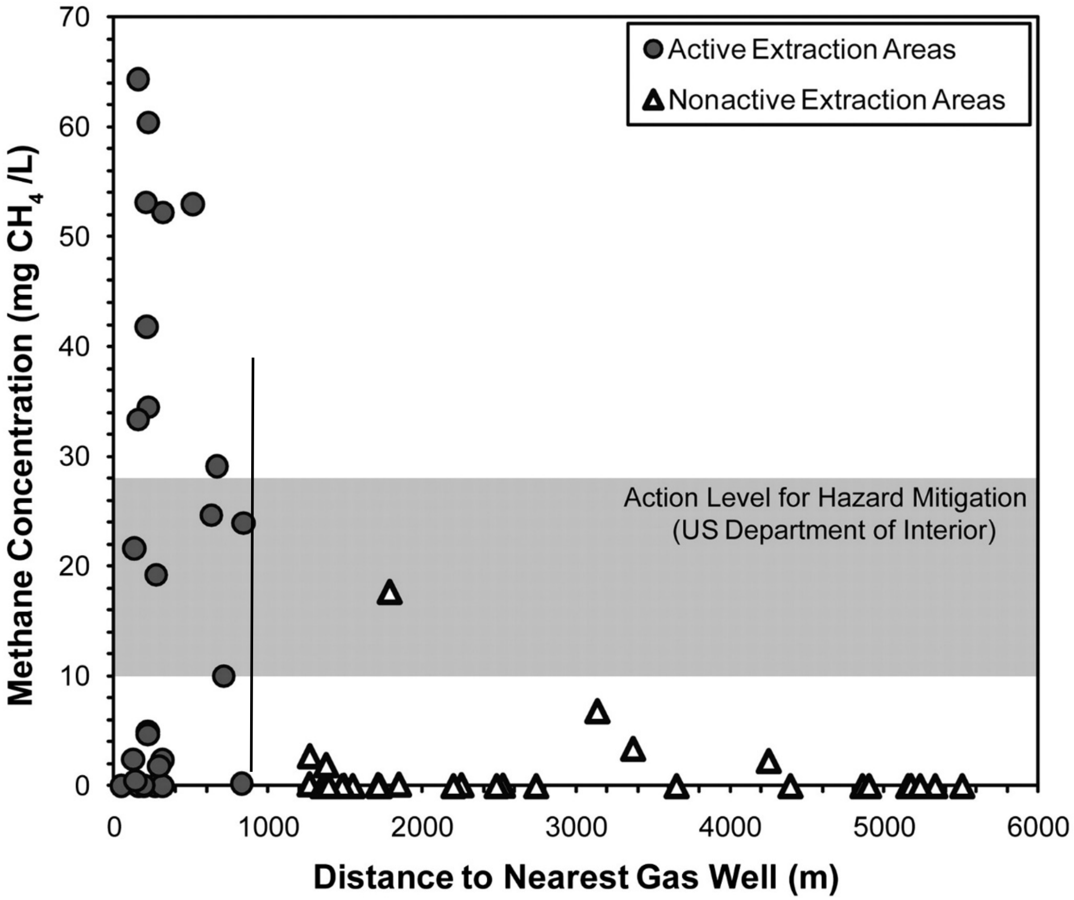
Osborn, Stephen G., et al. "Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing." proceedings of the National Academy of Sciences 108.20 (2011): 8172-8176.

MLA Siegel, Donald I., et al. "Methane Concentrations in Water Wells Unrelated to Proximity to Existing Oil and Gas Wells in Northeastern Pennsylvania." Environmental science & technology 49.7 (2015): 4106-4112.

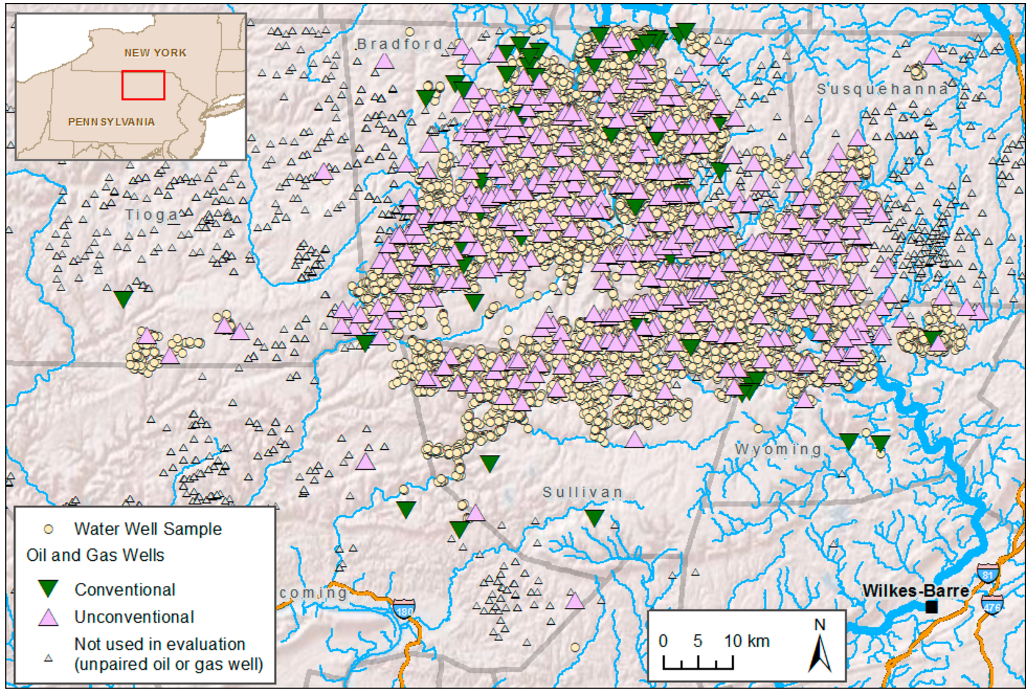
"Our results show evidence for methane contamination of shallow drinking-water systems in at least three areas of the region and suggest **important environmental risks** accompanying shale-gas exploration worldwide." (Osborn et al., 2011)

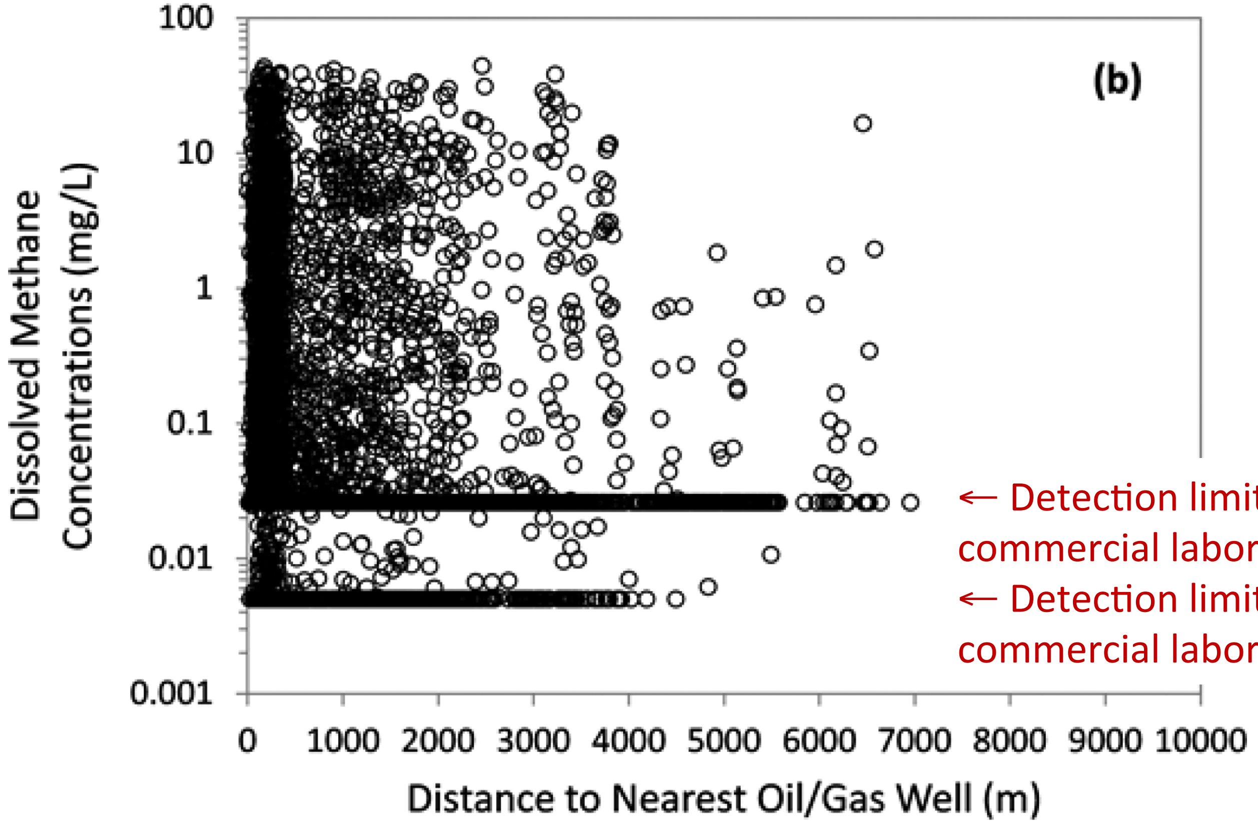


Analyzed groundwater from **68** private water wells



"In contrast to prior findings, we found <u>no statistically significant relationship</u> between dissolved methane concentrations in groundwater from domestic water wells and proximity to pre-existing oil or gas wells." (Siegel et al., 2015)





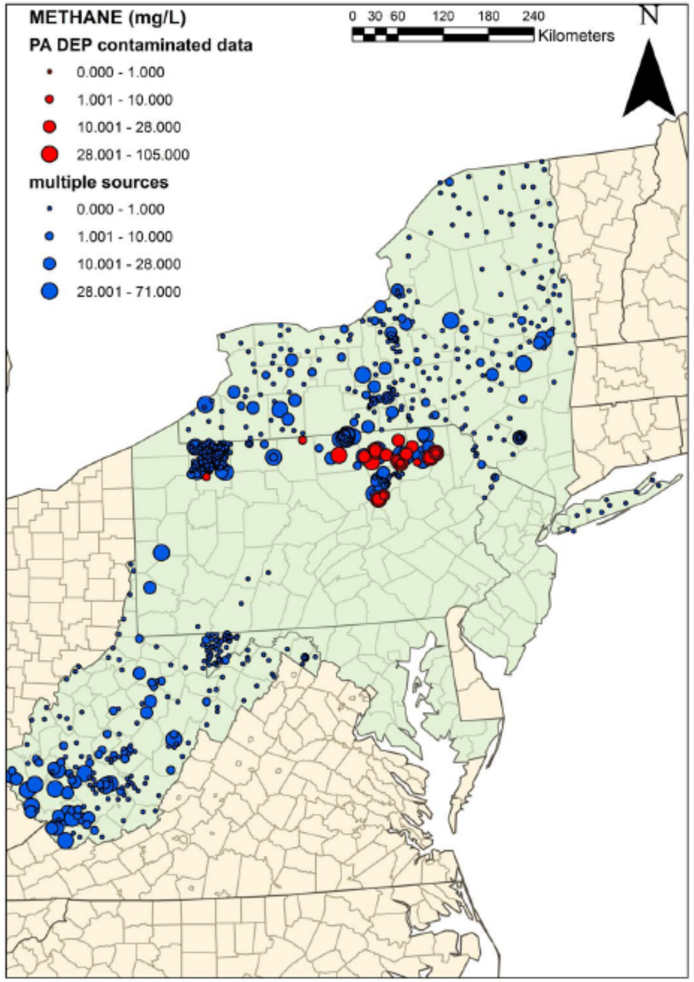
← Detection limit for some commercial laboratories

← Detection limit for some commercial laboratories

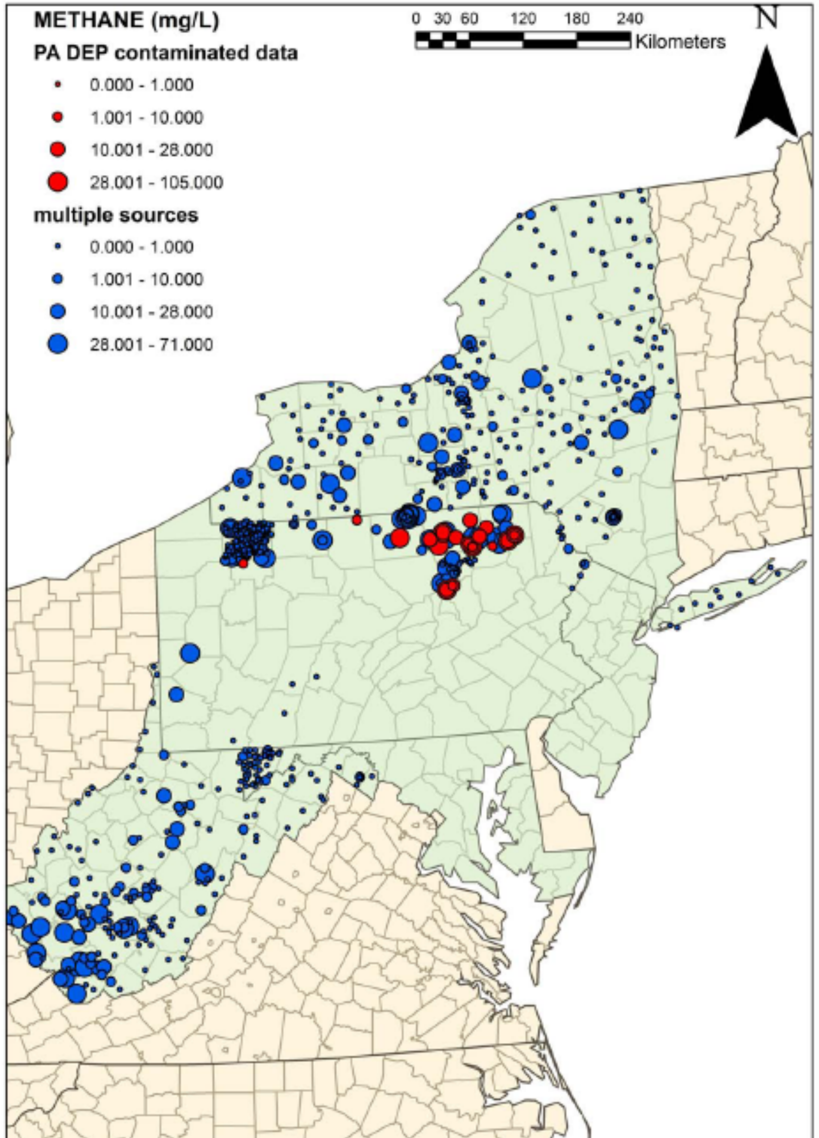
- Analyzed **11,309** data samples
- 661 pre-existing gas wells (wells exist before samples are taken)
- 639 gas wells, 22 oil wells
- 56 conventional wells, 605 unconventional wells
- Data from Chesapeake Energy in Bradford county

Possible reasons why Osborn et al. saw a trend with distance and Siegel et al. did not

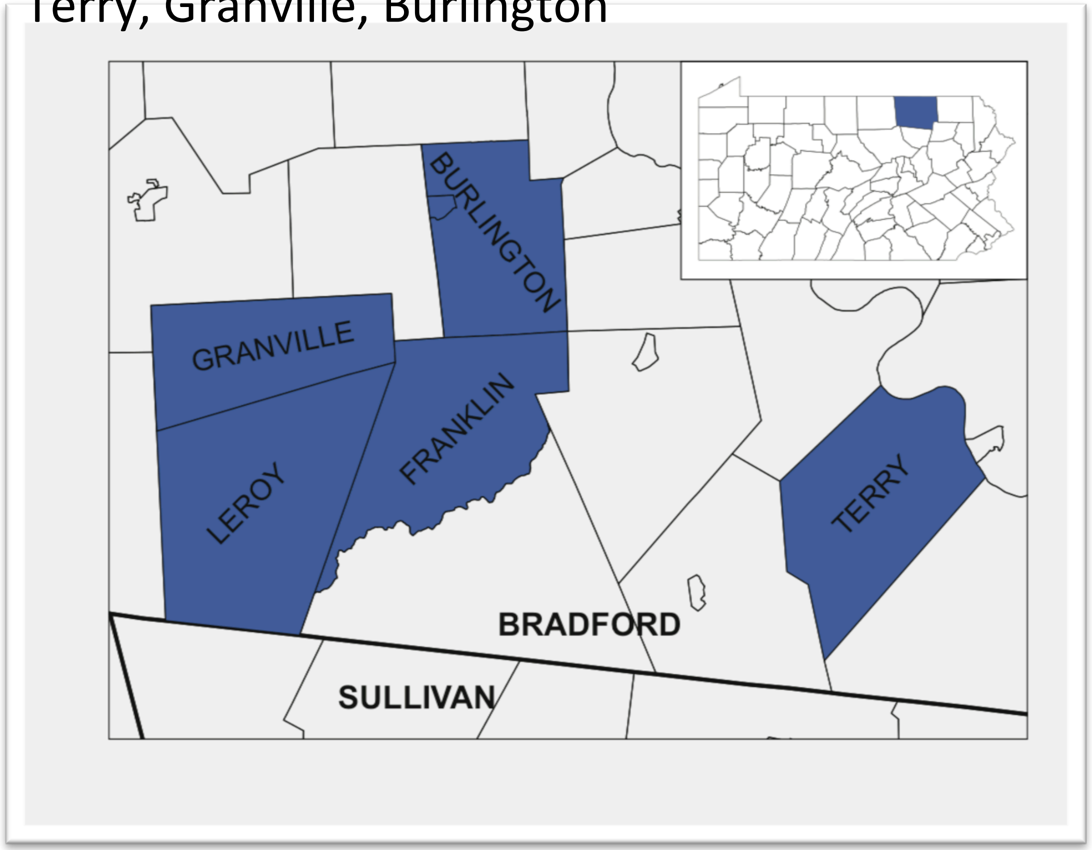
- Osborn's data set was too small (68 samples), leading to an inaccurate conclusion
- A large dataset in Siegel (11,309 samples) will show the average result (which is that there are few problems): problems can only be distinguished when small problematic regions are analyzed
- Some areas have problems and some do not (Osborn's regional data coverage was not identical to that of Siegel and in fact included data from Dimock PA where some acknowledged problems occurred)



Our dataset: DEP data from 5 townships

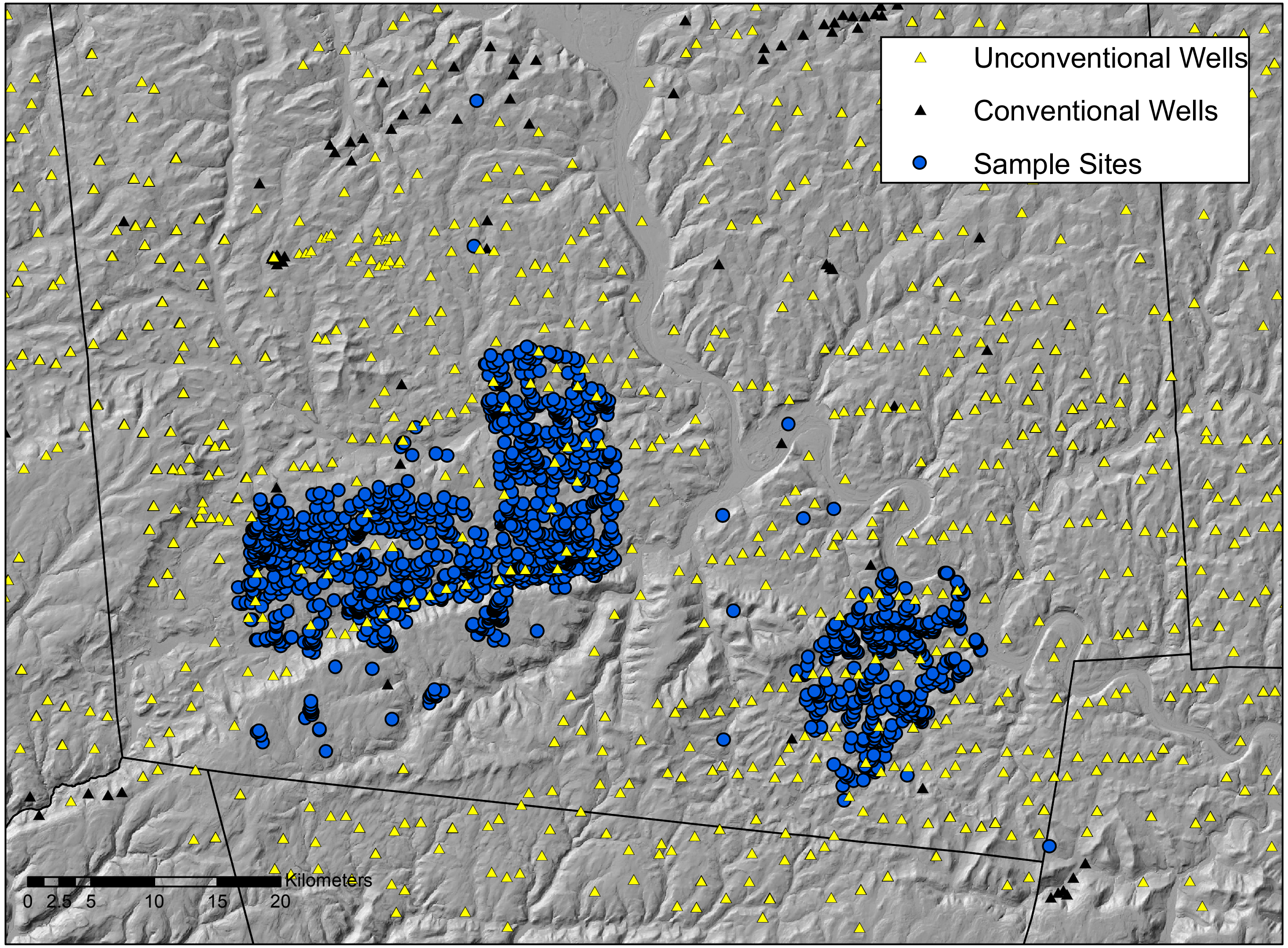


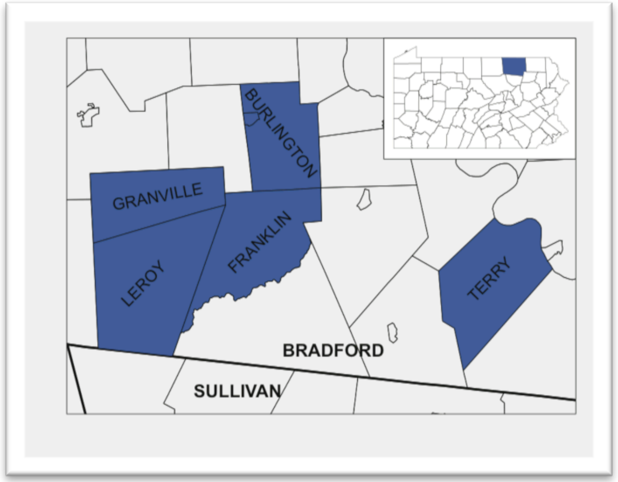
Five townships (with known problems): Franklin, Leroy, Terry, Granville, Burlington



All the data have been uploaded to ShaleNetwork database, accessible from Hydrodesktop

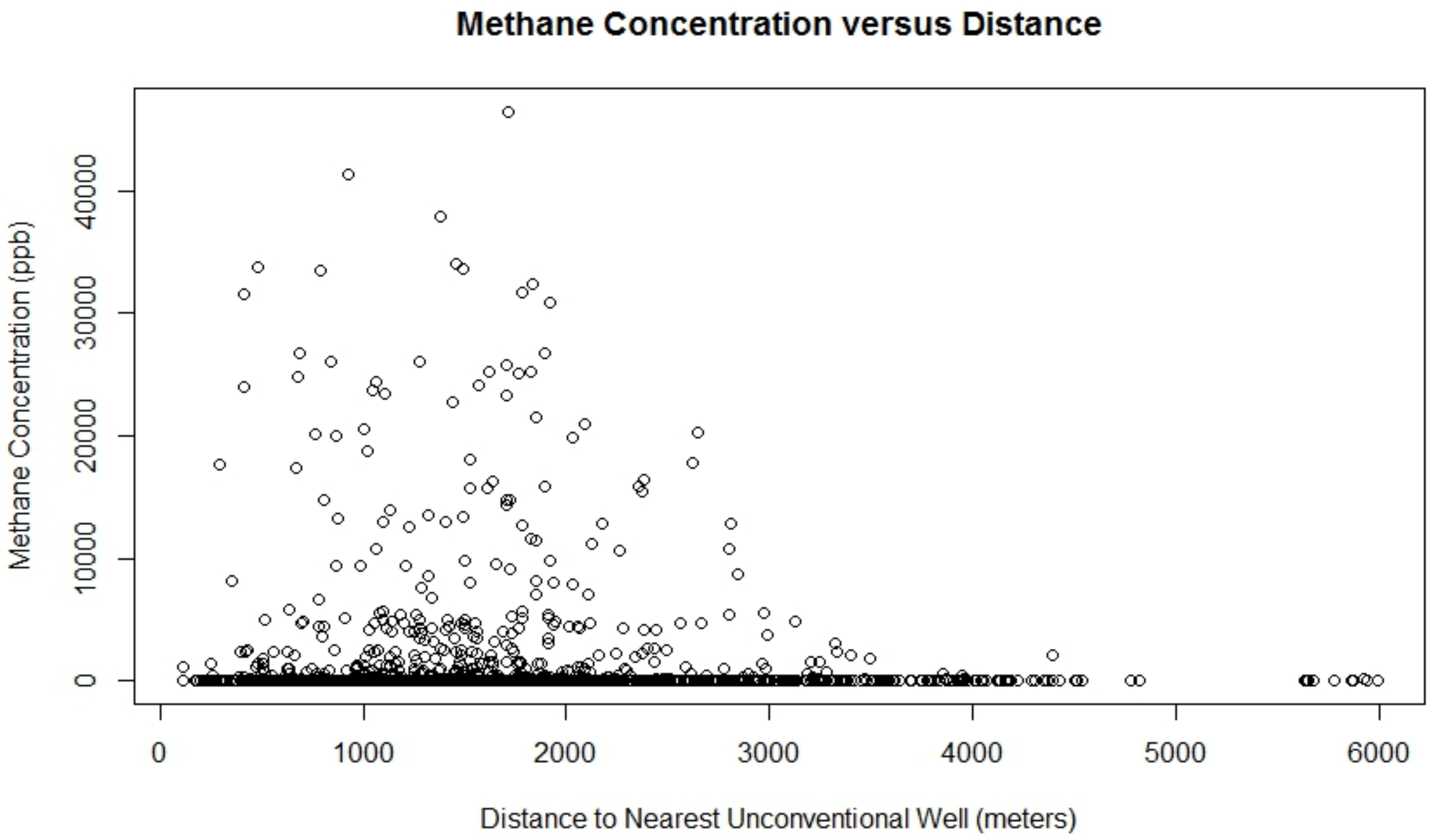
Our dataset: DEP data from 5 townships





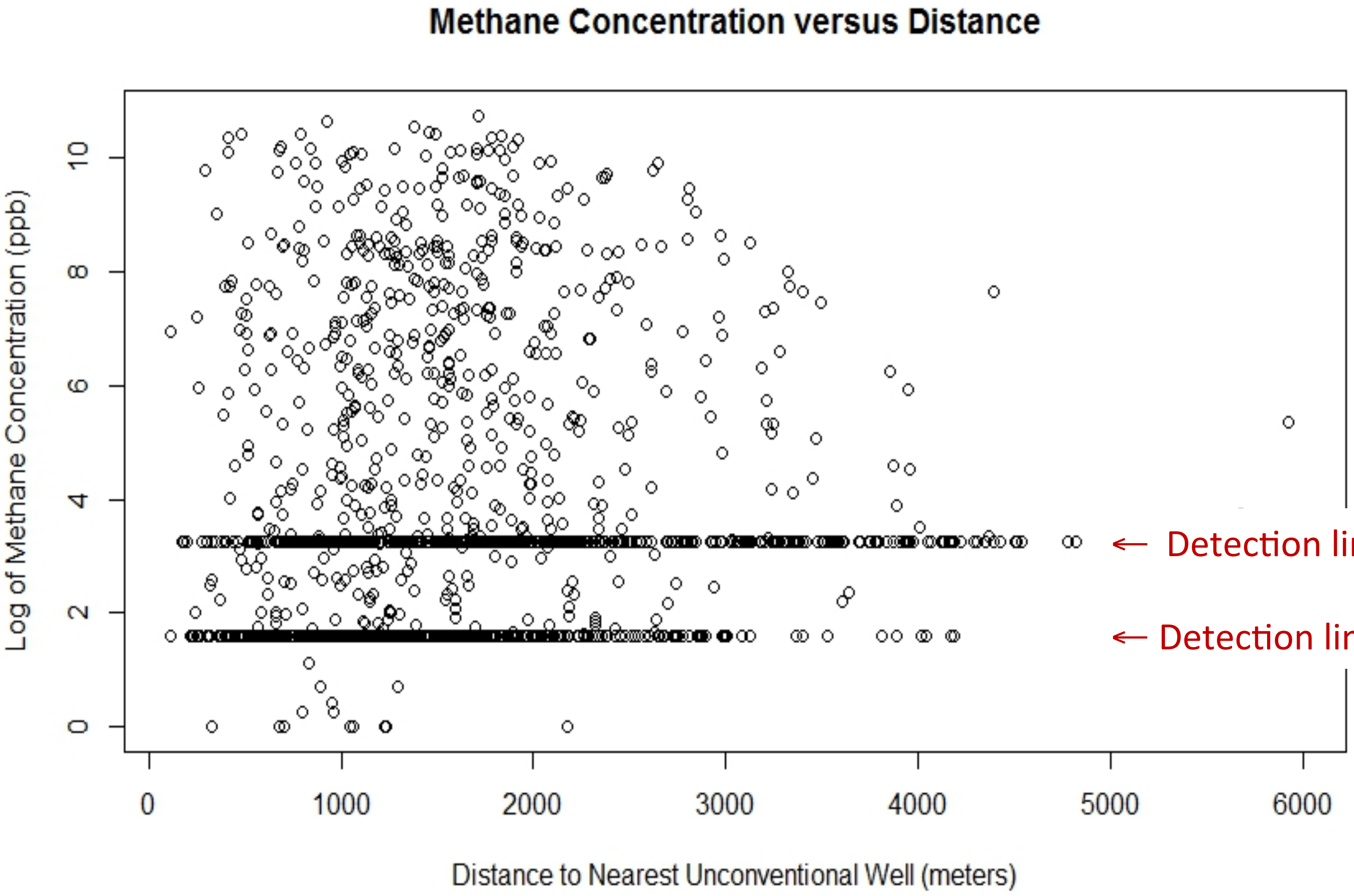
- 1643 groundwater sources
  - 1525 wells, 118 springs
  - Dec 20, 2010 - Nov 23, 2012
  - ~33 analytes
- gas wells
  - 1151 unconventional wells (spud before 3/21/2013)
  - 68 conventional wells (spud before 5/7/2009)

Methane Concentration vs. Distance to Nearest Gas Well



Methane Concentration vs. Distance to Nearest Gas Well (log scale)

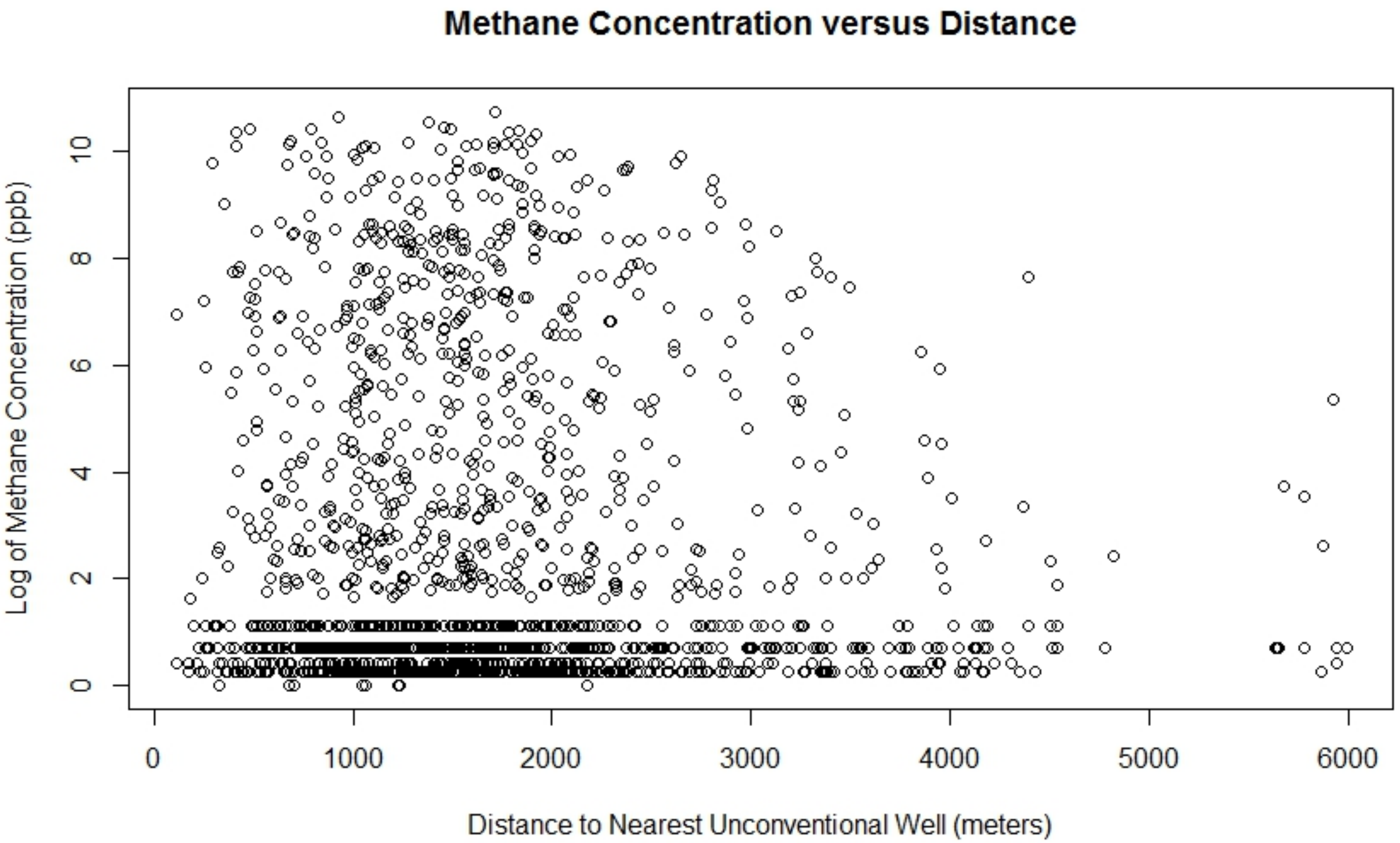
We cannot calculate averages for these data because **71%** are reported as beneath detection. We use a statistical technique called BOOTSTRAPPING to estimate data beneath detection so that it has the same statistical character as the rest of the data



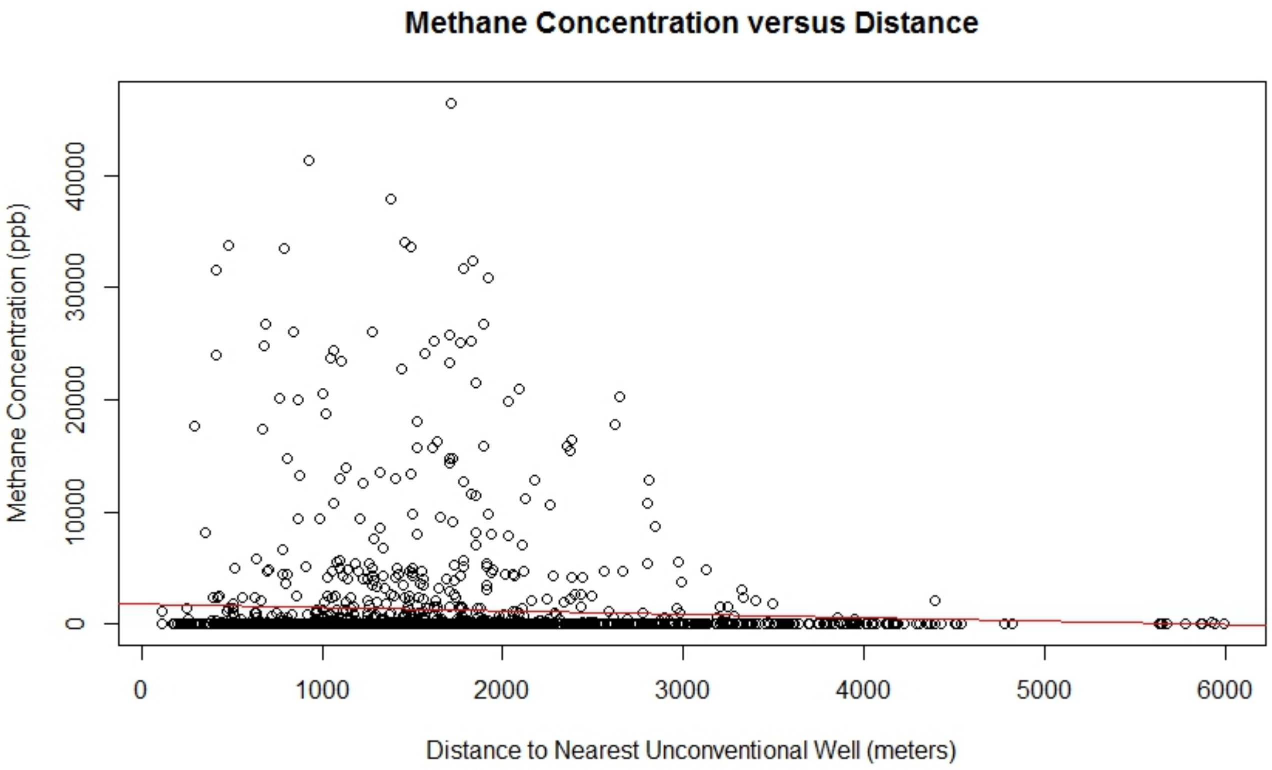
← Detection limit for some commercial laboratories

← Detection limit for some commercial laboratories

Methane Concentration vs. Distance to Nearest Gas Well (log scale, bootstrapped)



Methane Concentration vs. Distance to Nearest Gas Well (bootstrapped): Statistical Correlation



Pearson's Correlation	-0.061* p-value:0.01147
Spearman's Correlation	-0.018 P-value: 0.4452
Kendall's rank Correlation	-0.012 P-value: 0.443
Regression's Analysis	-0.2968* P-value: 0.01147

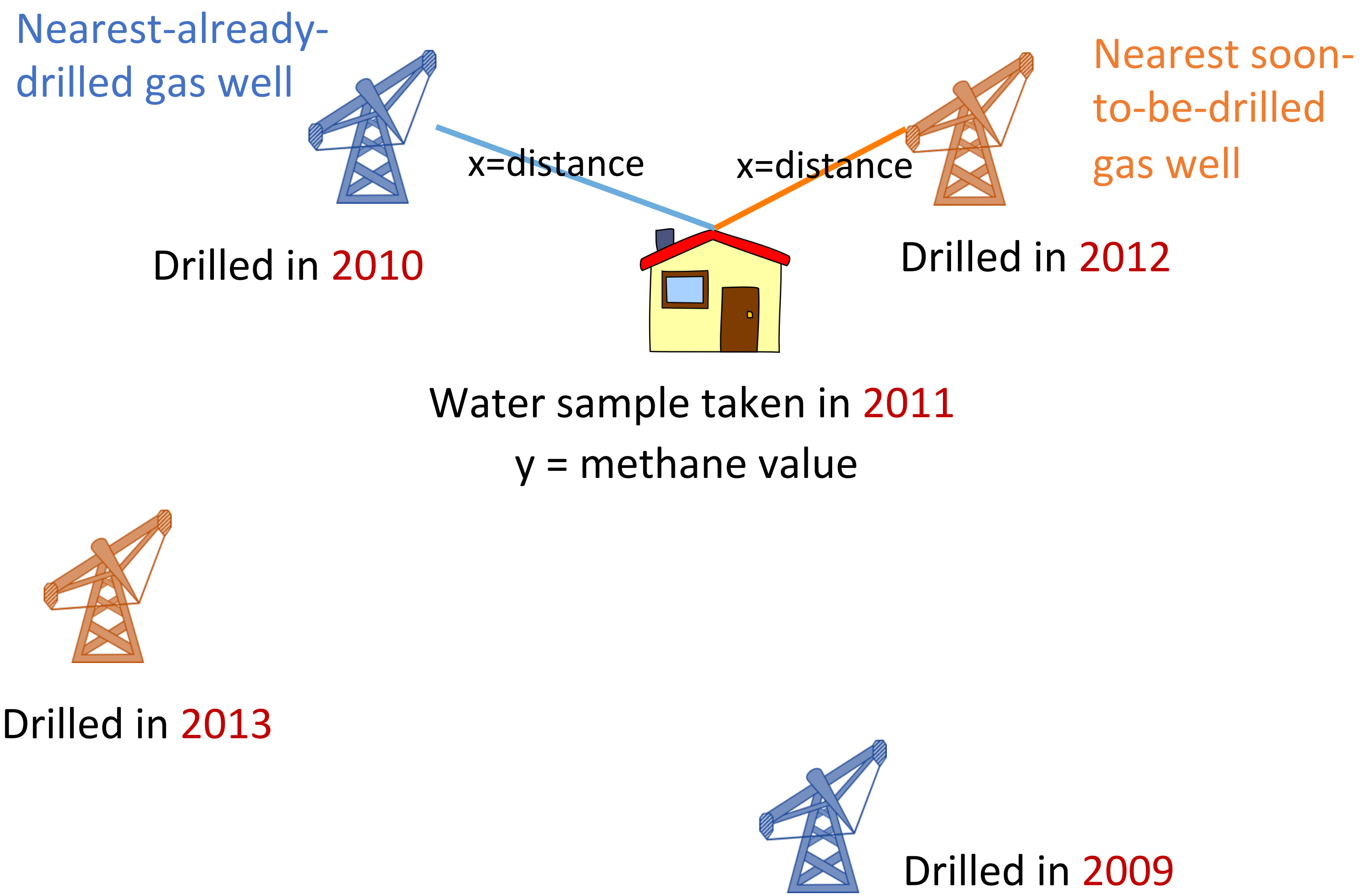
Conclusion: We do see **a weak but statistically significant** increase in methane concentration closer to shale gas wells (within 3 km especially) and this could be caused the shale gas development.

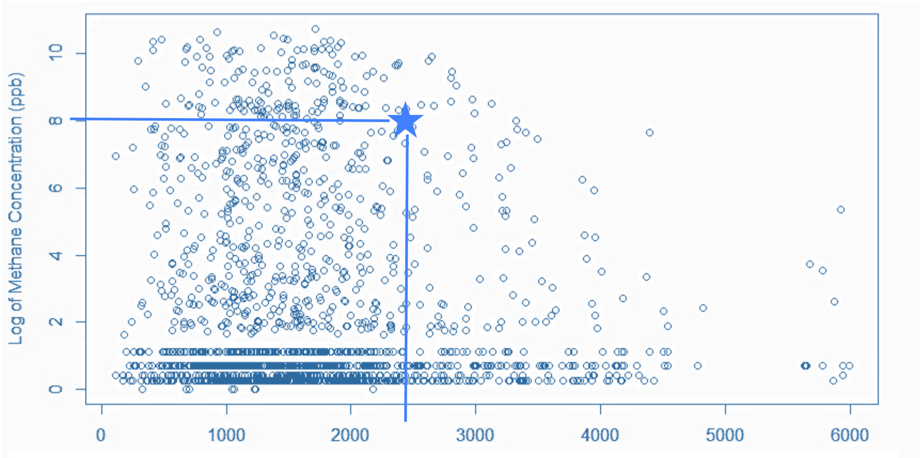
By law, gas companies in PA are presumed responsible for impacts on water if a water supply within 0.762 km (2500 ft) is impacted after unconventional gas development as compared to before, as long as a geological connection or explanation is likely.

Is this caused by shale gas well development, and if so, what is the explanation?

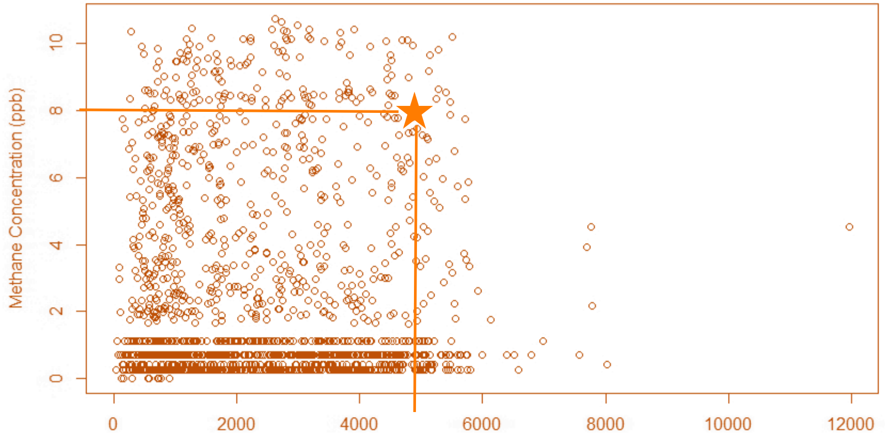
- Hypothesis 1: The significant weak correlation between methane and distance is due to some feature related to local geology, that is identical to the feature of the geology that leads companies to drill in certain locations. (Correlation to a third, underlying variable instead of causation)
- Test: Does the correlation exist before drilling?

Test: Does the correlation exist before drilling?



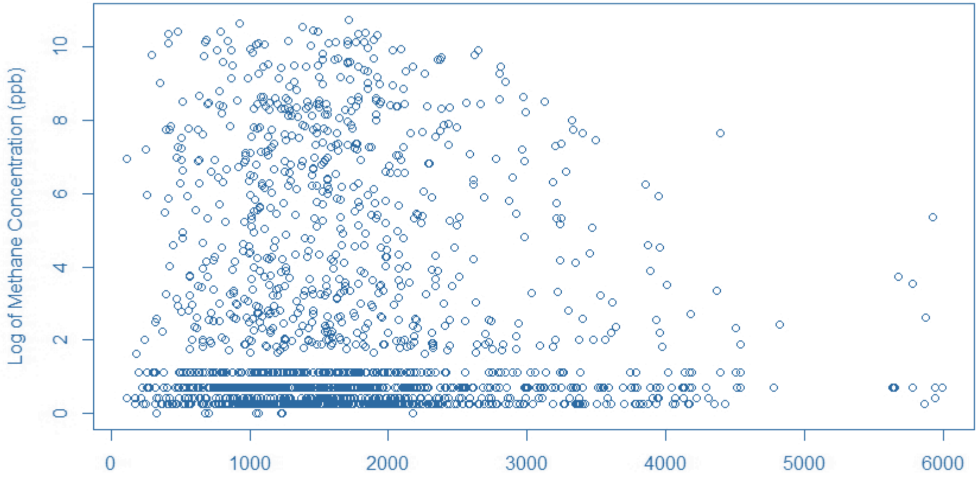


Distance to nearest **already-drilled** gas well

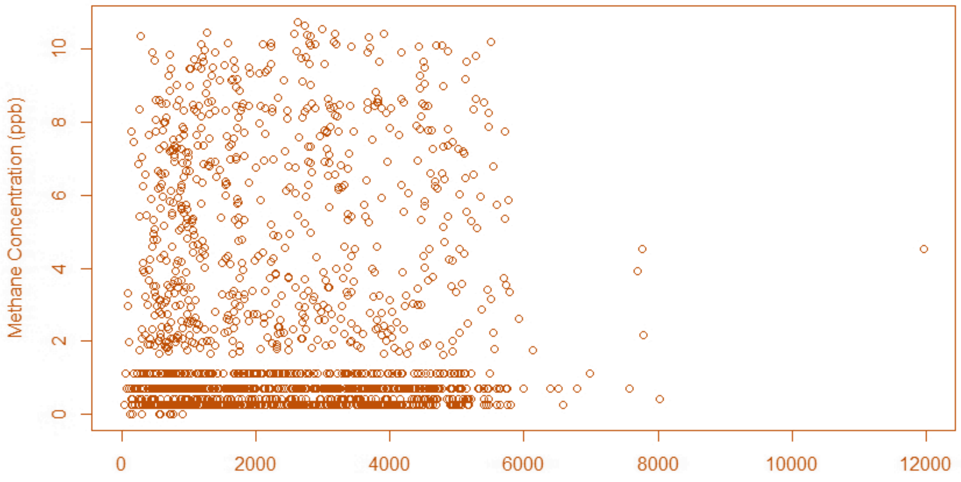


Distance to nearest **soon-to-be-drilled** gas well

Test: Does the correlation exist before drilling? Compare already-drilled with soon-to-be-drilled



Already-drilled



Soon-to-be-drilled

	Already-drilled	Soon-to-be-drilled
Pearson's Correlation	-0.061* p-value:0.0115	0.0313 p-value:0.19
Spearman's Correlation	-0.018 P-value: 0.4452	-0.013 P-value: 0.58
Kendall's rank Correlation	-0.012 P-value: 0.443	-0.008 P-value: 0.59
Regression Analysis	-0.2968* P-value: 0.0115	0.0894 P-value: 0.196

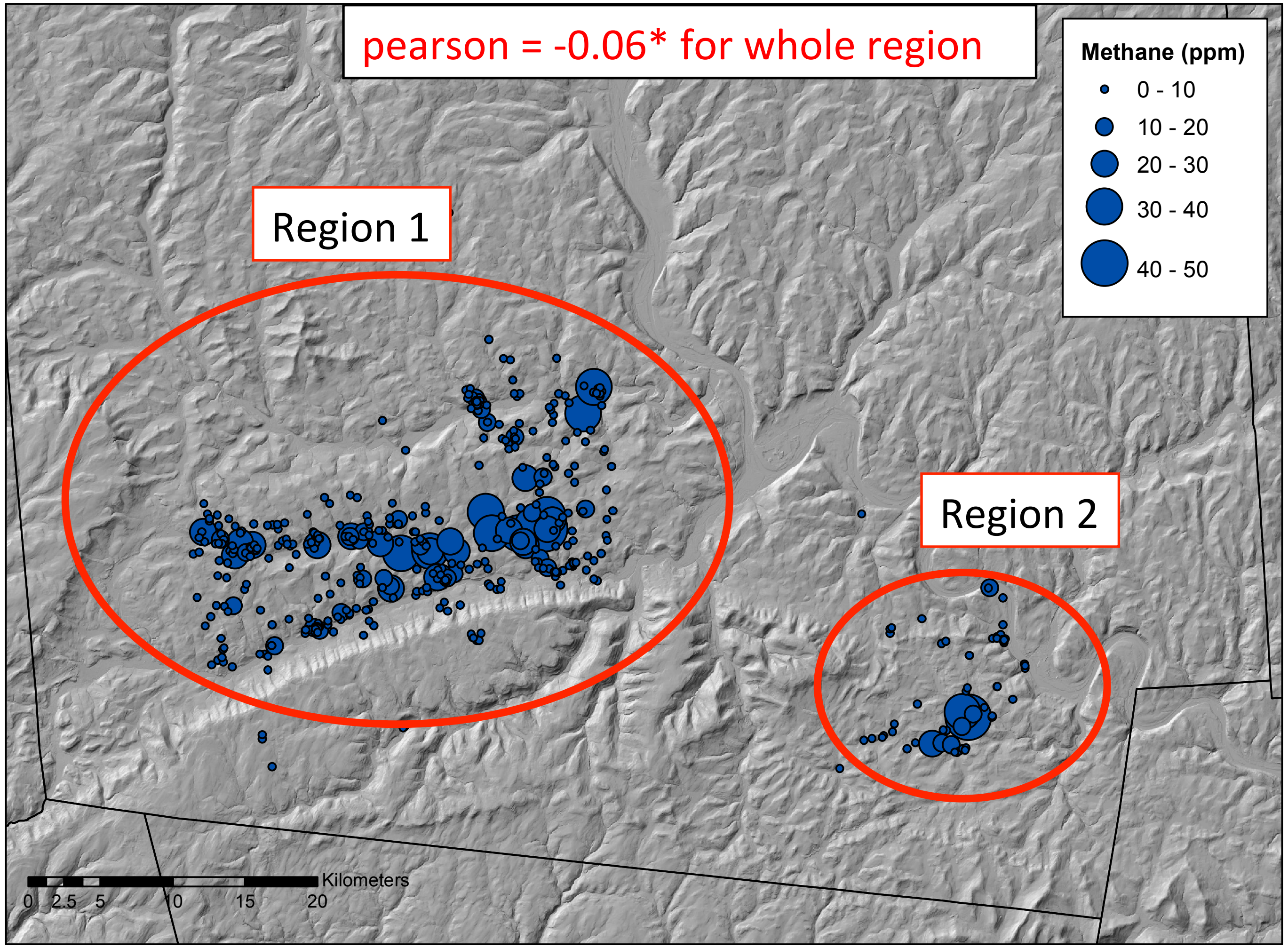
Conclusion: We do *not* see a statistically significant correlation in soon-to-be-drilled wells: where they chose development sites does not cause spatial variation in methane.

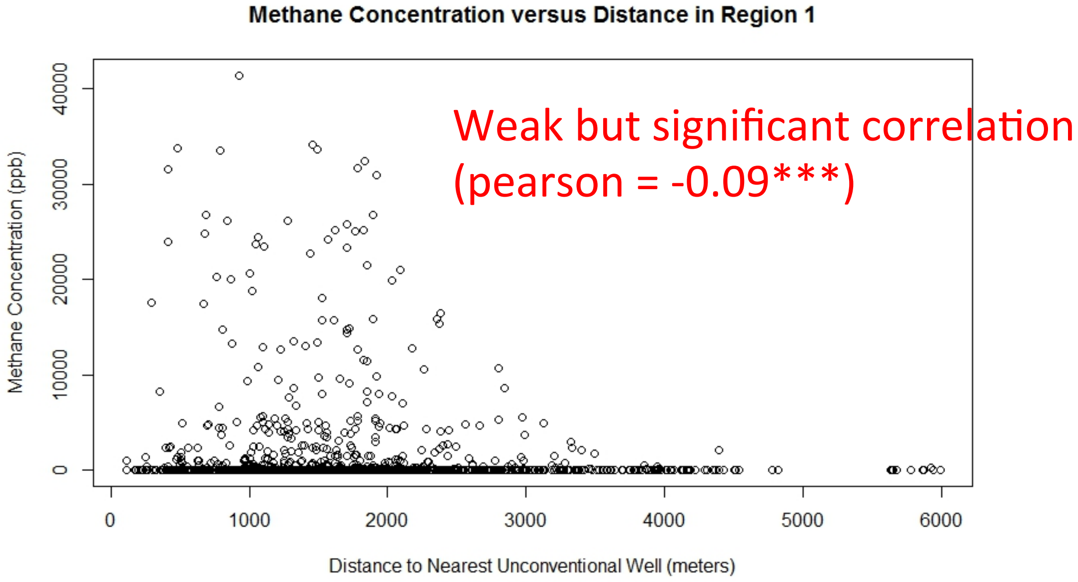
Bootstrapped data

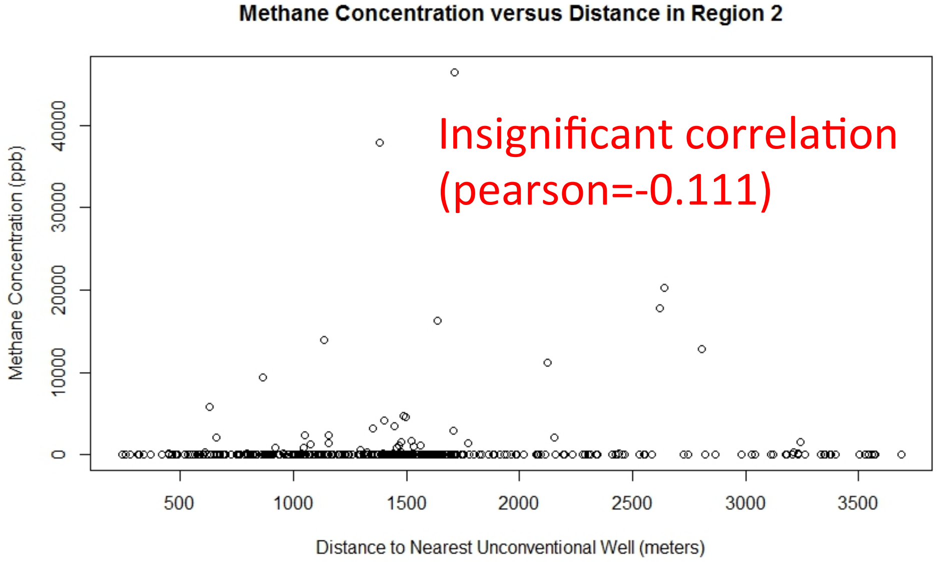
Is this caused by the shale gas well development, and if so, what is the explanation?

- Hypothesis 2: The correlation is only significant in certain smaller sub-regions of the overall area.
- Test: Examine the correlation at smaller regions separately.

Analysis by sub-region



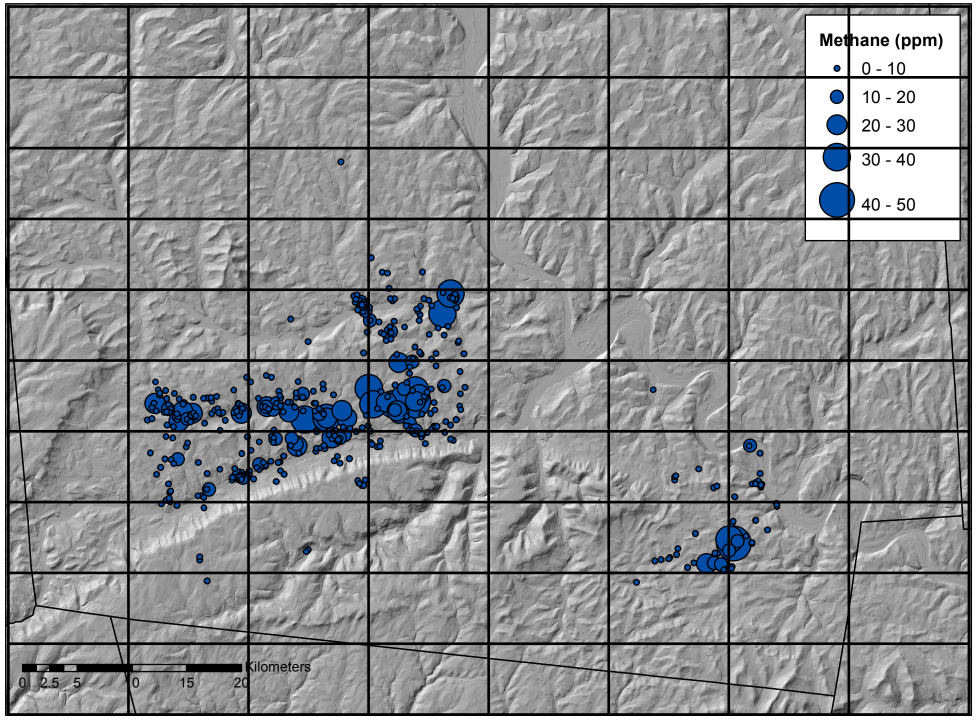


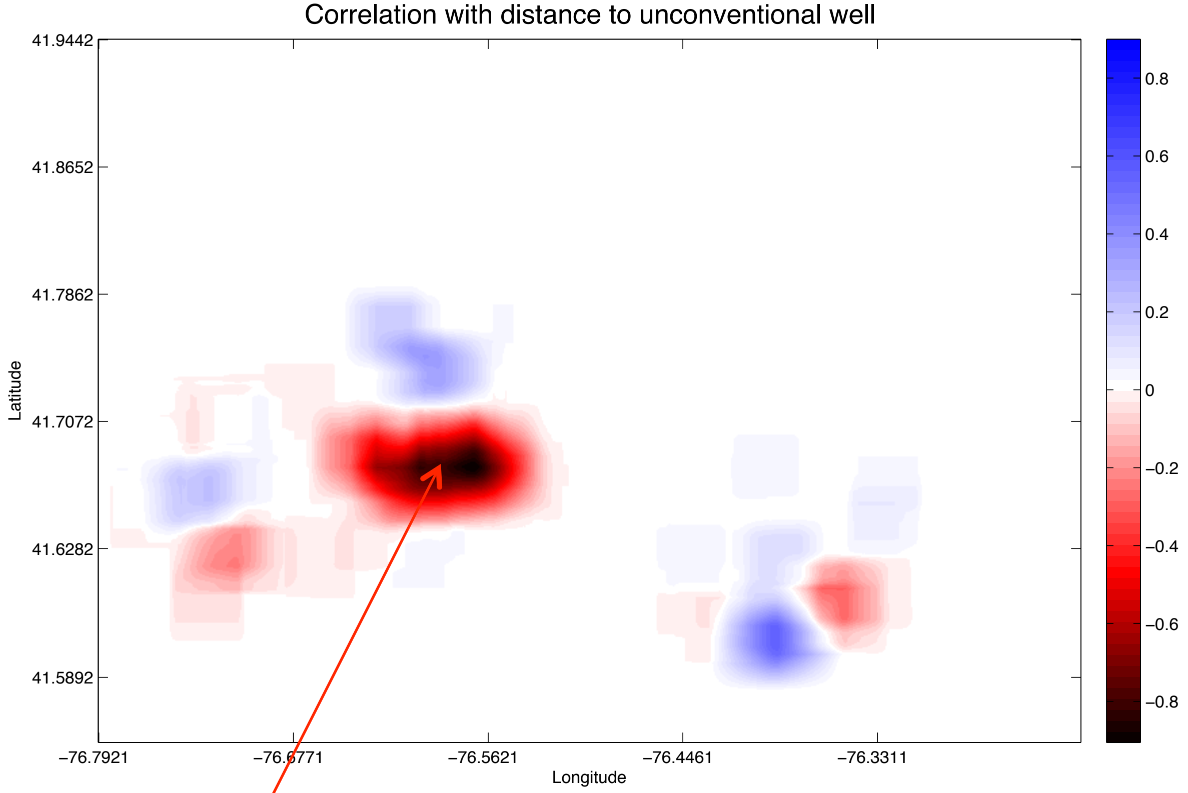


Region Analyses

Red: Significant negative correlation (i.e. water near gas wells has more methane)

Blue: Significant positive correlation (i.e. water near gas wells has less methane)





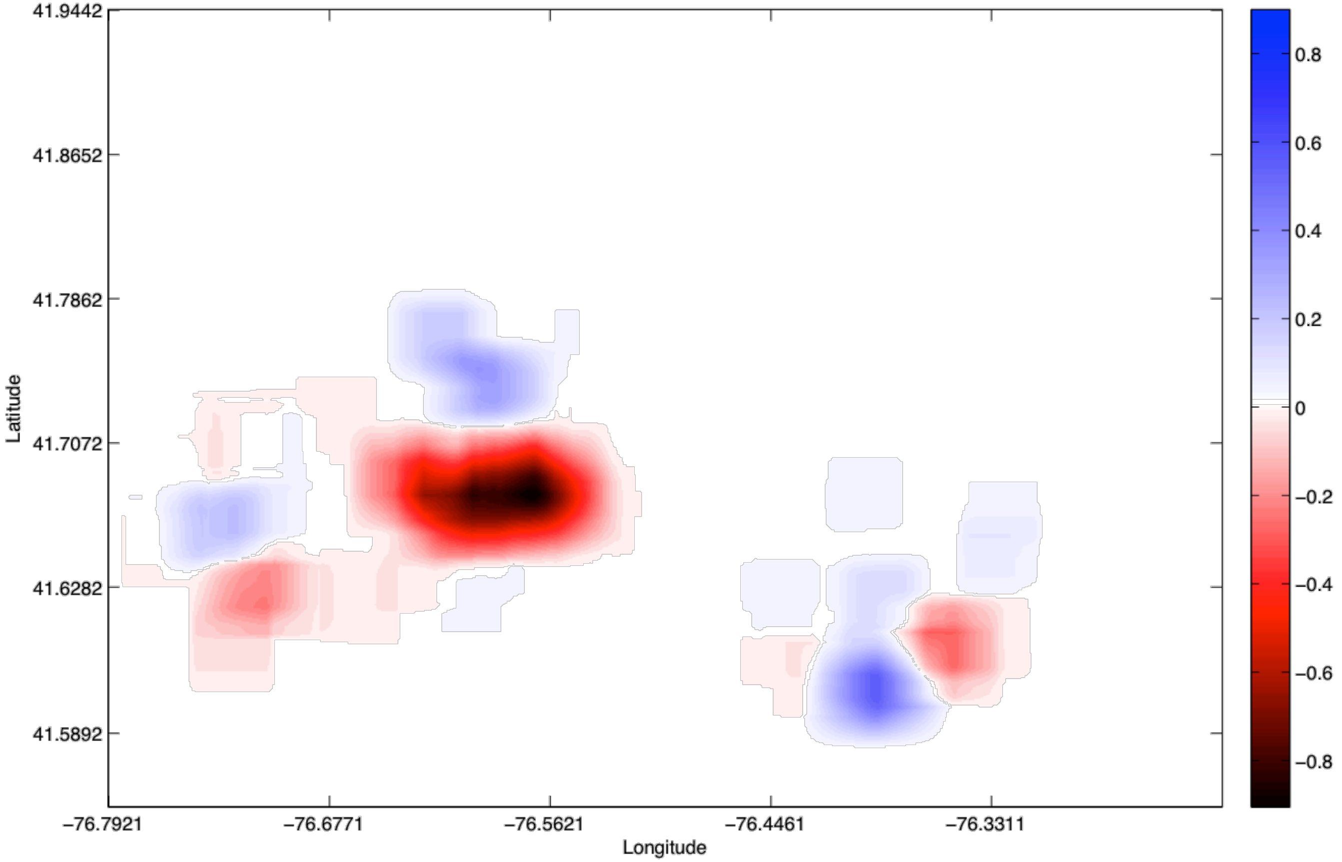
Conclusion:

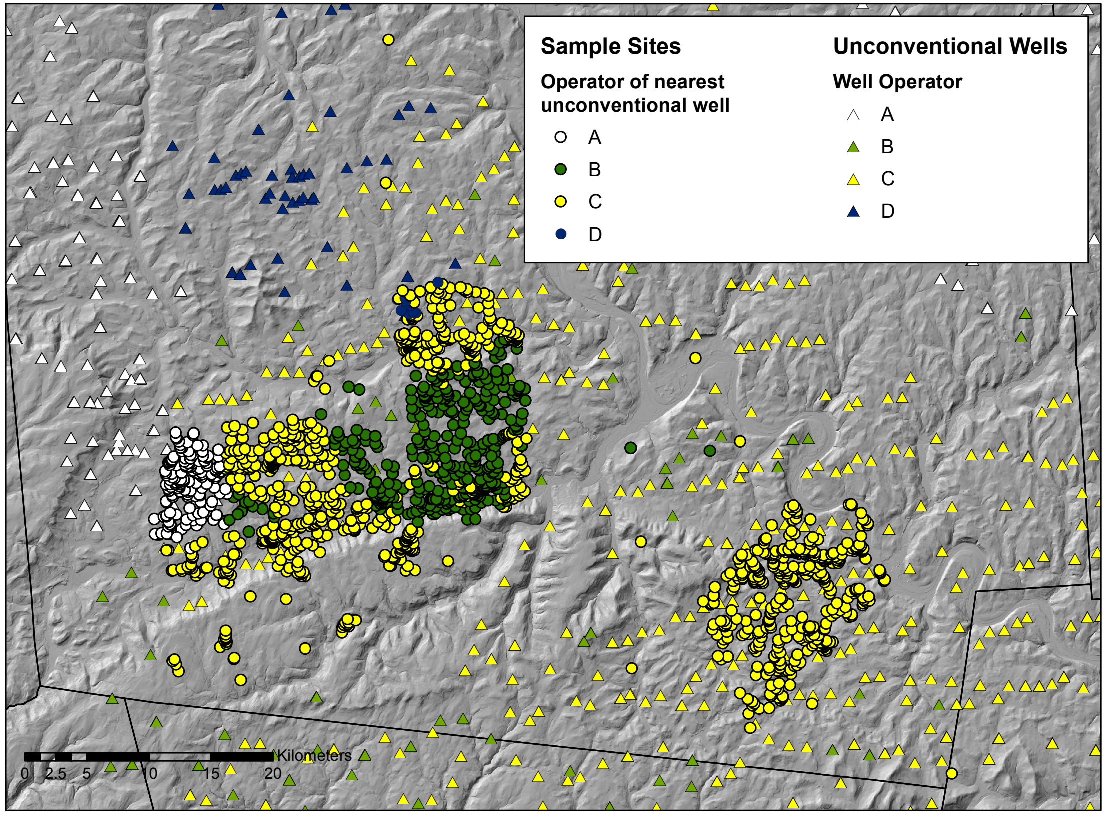
Some regions (red) show that waters near gas wells tend to have slightly more methane.

Is this caused by the shale gas well development, and if so, what is the explanation?

- Hypothesis 3: The very small increase in methane concentrations observed near shale gas wells is due to specific company practices that vary because the identity of the company that owns leasing rights varies by region.
- Test: how do the high methane concentrations vary with spatial position and can we see any correlation between high methane and company identity that might explain this variation?

Methane Concentrations and Companies



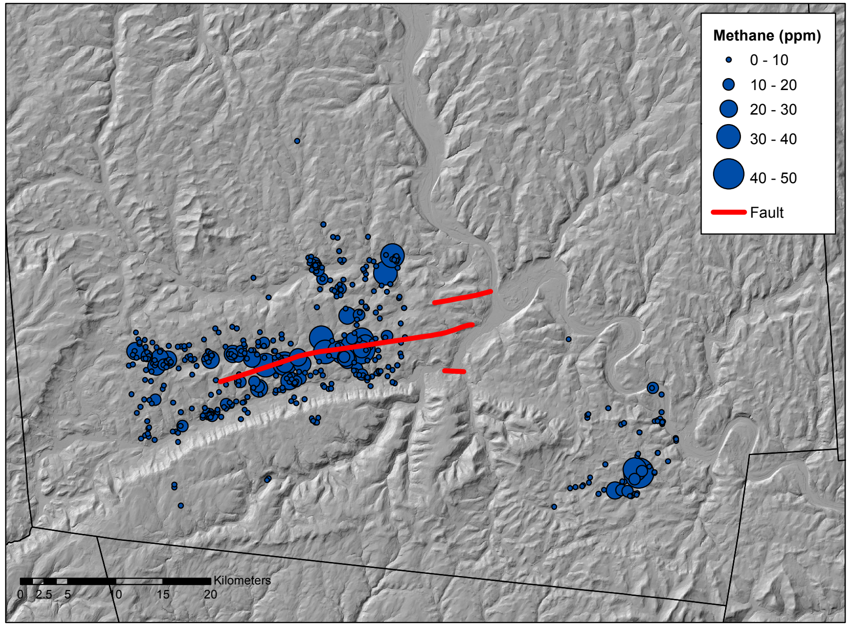


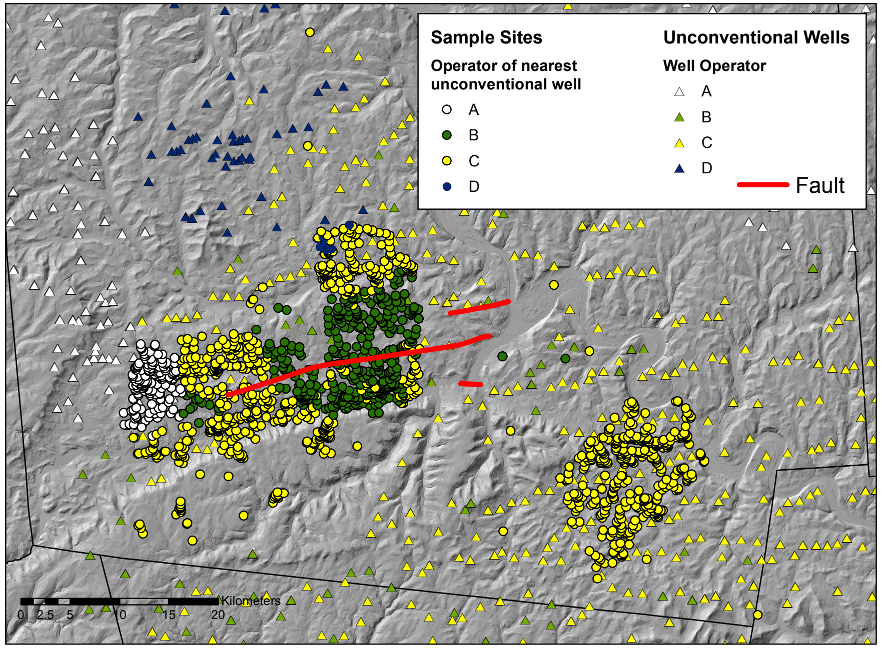
Conclusion: One company is more associated with problems than the other 3 companies

Is this caused by the shale gas well development, and if so, what is the explanation?

- Hypothesis 4: The very small increase in methane concentrations observed near shale gas wells is due to local geology in certain sub-regions and this explains the location of problems rather than company practice
- Test: how do the high methane concentrations vary with spatial position and can we see any geologic reason for this variation?

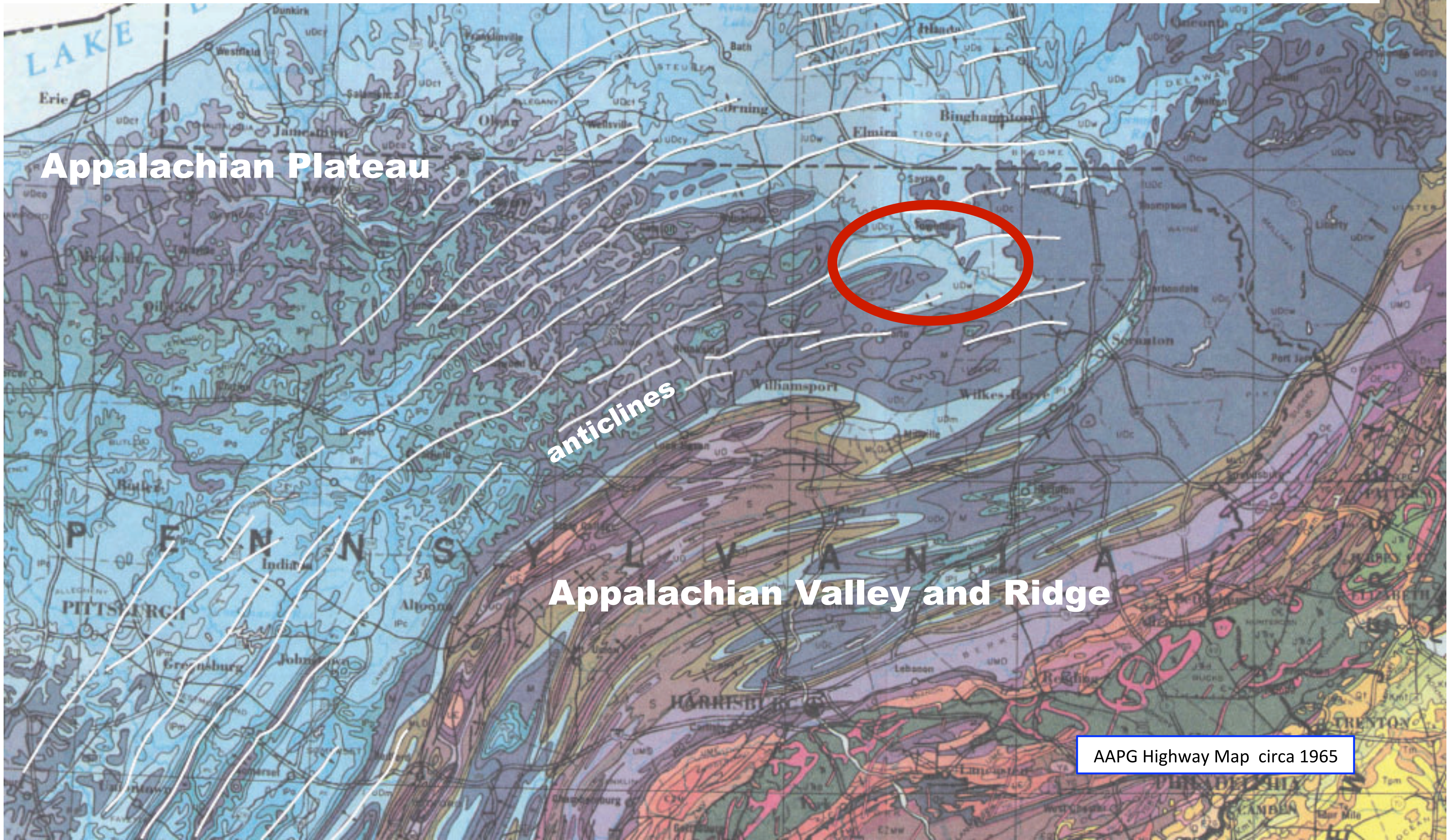
How do the high methane concentrations vary with spatial position and can we see any geologic reason for this variation?



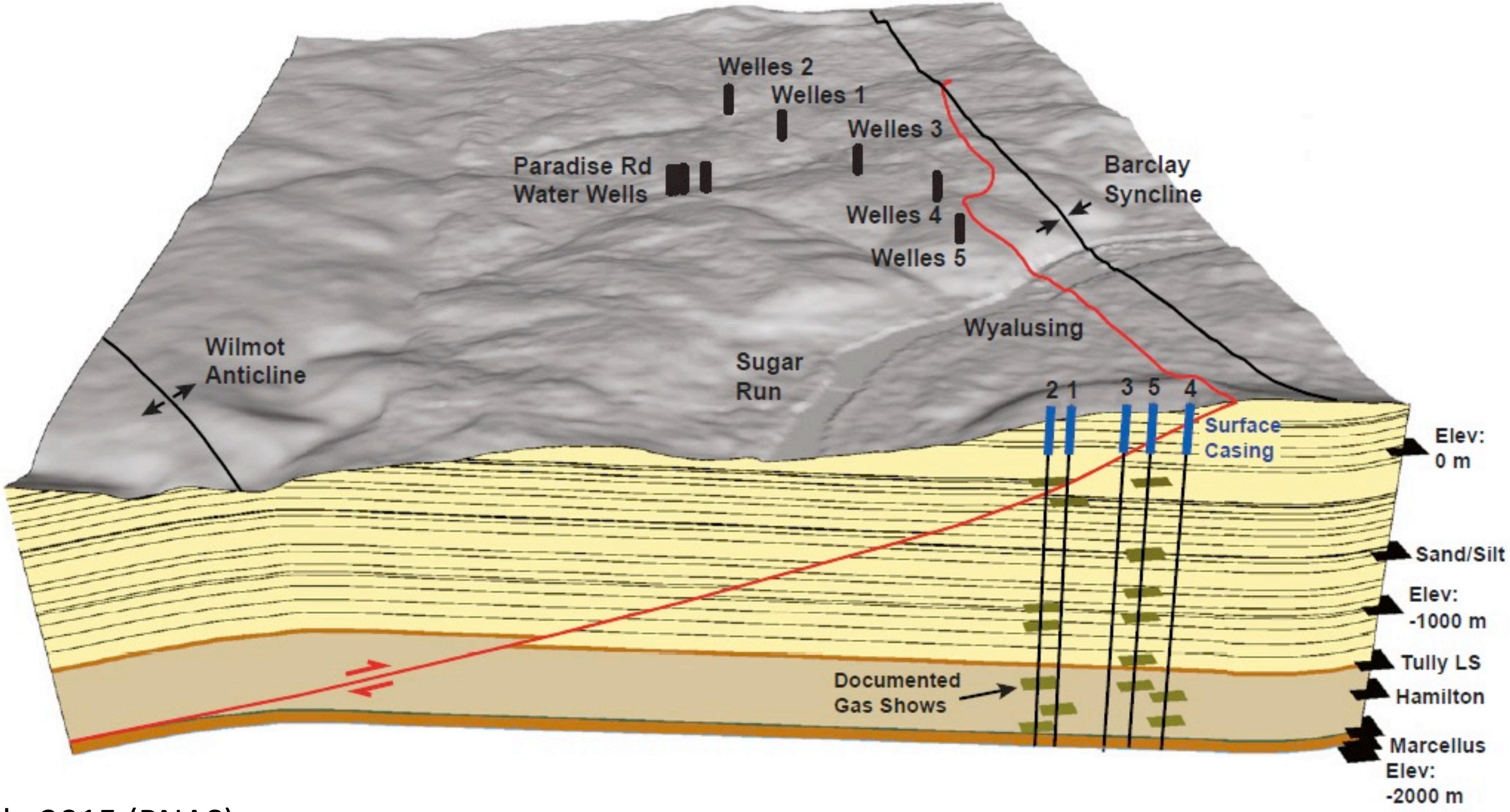


Conclusion: Many of the high methane values in ground waters are near a large, mapped fault. The one company associated with high methane values may have leased land in this area and developed it with best practices – and yet this part of the landscape has a greater tendency for methane to seep into groundwaters because of the presence of the fault.

Many large faults in Pennsylvania have been mapped and related to anticlines/synclines



Schematic diagram of relationship of fault trace on land surface with underlying fault



Lewellyn et al., 2015 (PNAS)

Is this caused by the shale gas well development, and if so, what is the explanation?

- Hypothesis 5: Shale gas wells are drilled at locations with certain geological features – not only methane is correlated, but also other analytes
- Test: Examine the correlation with other analytes

Examine the correlation with other analytes

Distance ~ Barium + CaCo3 + Chloride + Ethane + Iron + MBAS + OilandGreaseHEM + pH + Propane + Selenium + Strontium + Sulfate + TDS

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

rank	Analyte	Coefficient	P	Signif.
1	Strontium	-1.38E-01	0.02017	*
2	TotalDissolvedSolids	-1.36E-01	0.07886	.
3	Barium	1.27E-01	0.06373	.
4	Chloride	1.24E-01	0.05583	.
5	Ethane	-1.01E-01	0.00338	**
6	pH	-9.96E-02	0.00442	**
7	Propane	8.39E-02	0.03658	*

rank	Analyte	Coefficient	P	Signif.
8	Iron	7.87E-02	0.00773	**
9	OilandGreaseHEM	7.82E-02	0.02818	*
10	Sulfate	7.09E-02	0.09602	.
11	BicarbonateAlkalinityasCaCO3	6.71E-02	0.1057	
12	MBAS	-5.63E-02	0.05214	.
13	Selenium	-4.40E-02	0.13002	

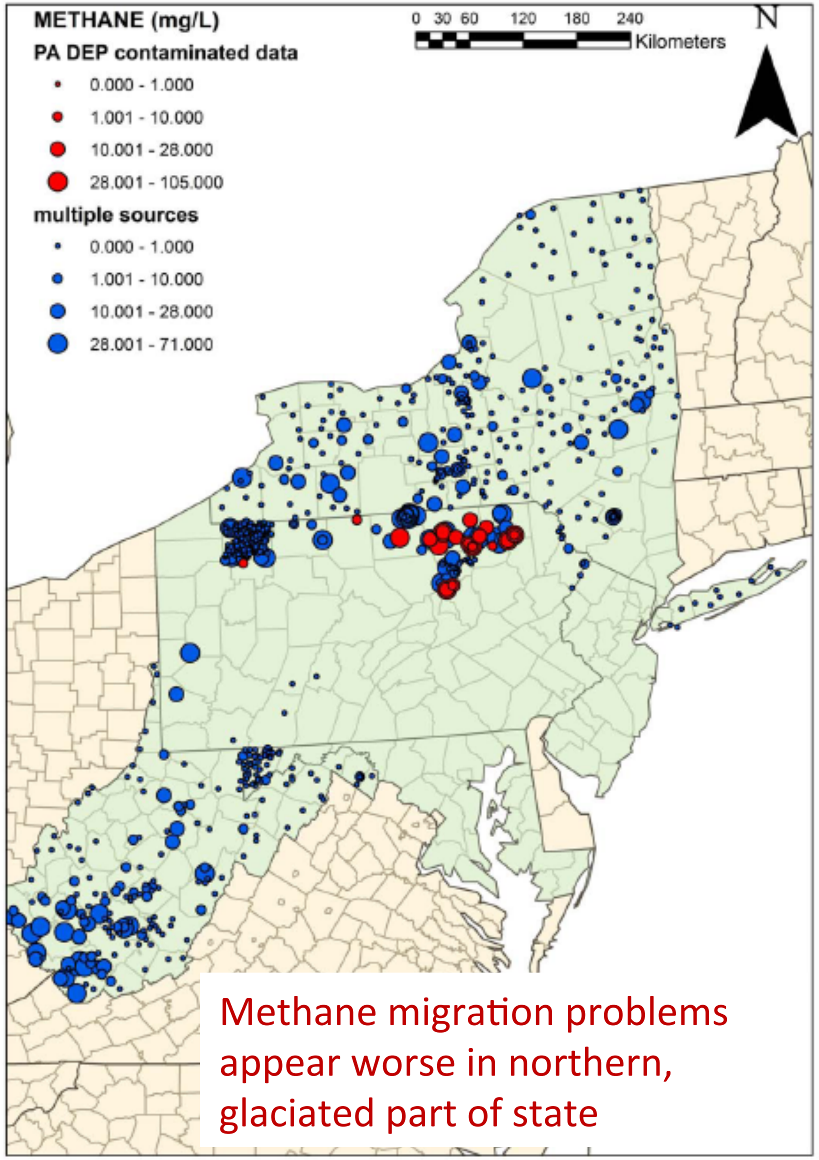
Conclusion: There are other analytes also correlate with distance to nearest already-drilled gas wells. These analytes may be introduced into the fault and may be indeed moving into ground water.

Conclusion

Acknowledgements: GE Gift Fund to Penn State (for J. Li, SLB, CY, FW, MG); NSF funding to SLB for Shale Network data; Jon Pollak of CUAHSI for help with data from HydroDesktop; data provision by PA DEP (working with Seth Palepko, Bill Kosmer).

- When a relatively large water quality dataset is considered (>400 datapoints) in a small area that has problems attributed to oil/gas activity, we see a **statistically significant but *very weak*** increase in methane concentration closer to already drilled shale gas wells
- This correlation ***could be caused*** by the shale gas activity or by a sampling artefact (because we don't have complete coverage for the region)
- This very weak increase in methane concentration is localized to **1000 km^2 subareas**
- The very weak increase in methane concentrations is **not** likely to be **related to pre-existing water quality** in our study area
- The very weak increase in methane concentration could be due to individual company practice but it is more likely due to the **presence of large faults** in the area
- There are **other analytes** also showing the correlation, which might be introduced by fault or company
- These type of analyses could help companies improve best practices in terms of locating gas wells to preserve water quality: however, analyses like this one require public release of water quality data and sample locations

Shale Network has been uploading methane data to our online database for sites where concentrations and locations have been made public. **Nonetheless, published methane concentration + location data in PA groundwaters are sparse**



Data from ShaleNetwork online database; figure made by PSU grad student Paul Grieve as of summer 2014.

Dots are scaled to the concentration level of methane in the ground water.

Blue circles were measured before high-volume hydraulic fracturing was implemented or represent measured values that were not deemed due to oil/gas activity by the PA DEP.

Red points: PA DEP attributed methane to oil/gas activity.

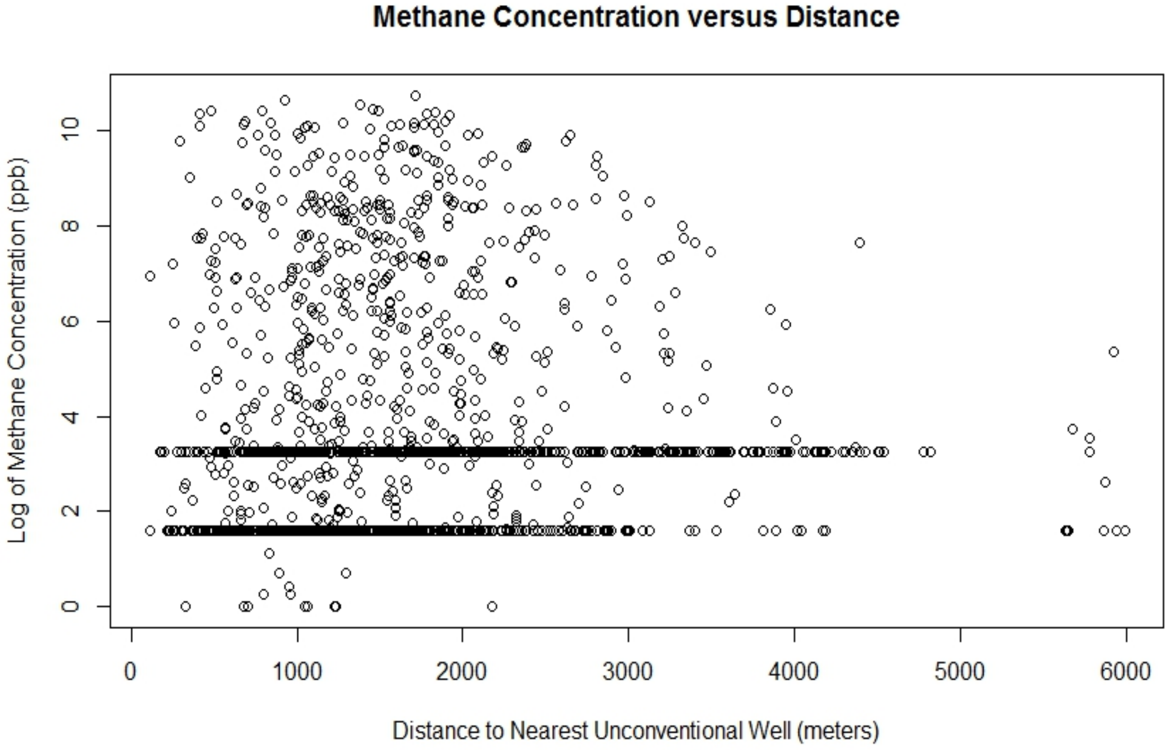
Data from Appalachia Hydrogeologic and Environmental Consulting, I., 2012; Breen et al., 2007; Kappell, W.M. and Nystrom, E.A., 2012; PA DEP data as summarized by L. Legere (Scranton Times-Tribune), see map of these determinations published with the letters of determination in collaboration with FracTracker (http://thetimes-tribune.com/news/gas-drilling-complaints-map-1.1490926?parentPage=2.2127); Moore, M.E. and Buckwalter, T.F., 1996; Pham, M.P. and Bolton, D.W., 2013; Sloto, R.A., 2013; White, J.S. and Mathes, M.V., 2006.

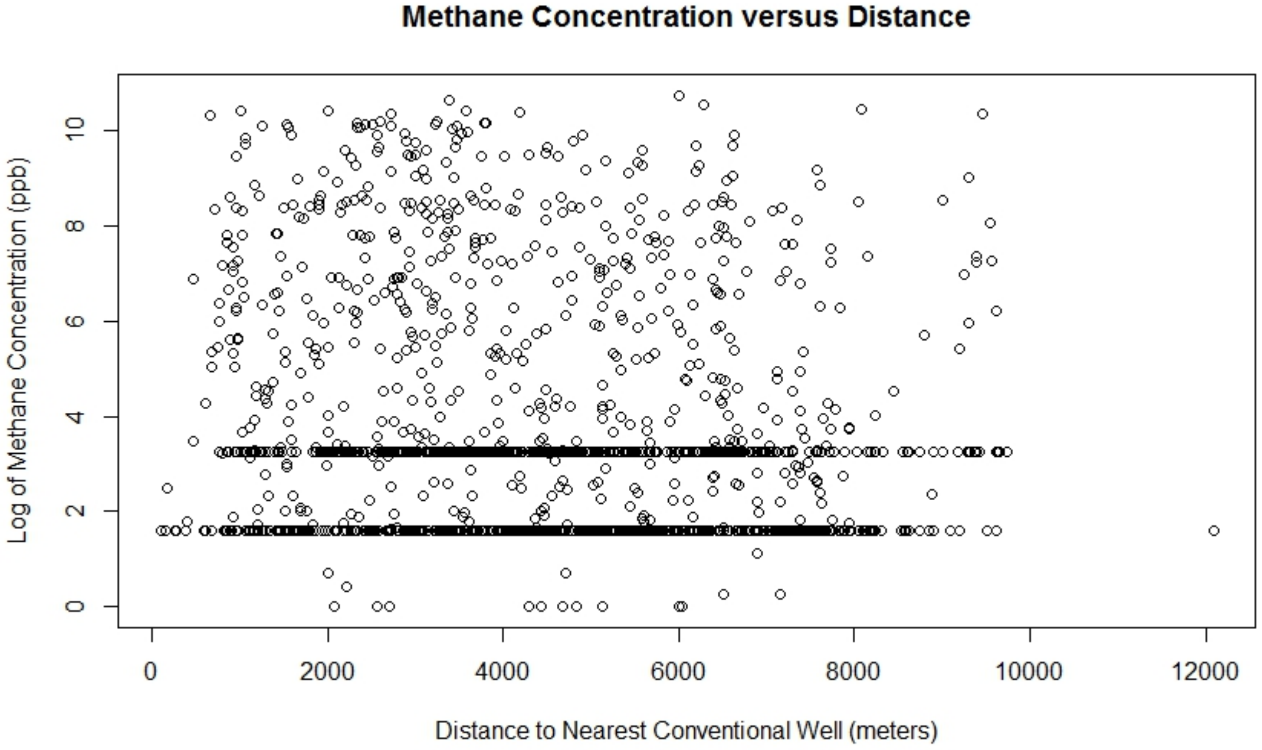
Brantley et al., 2014, Int. J. Coal Geology

Is this caused by the shale gas well development, and if so, what is the explanation?

- Hypothesis one: any kind of oil/gas well causes a perturbation that causes higher methane in water wells within 3 km
- Test: do we see a statistically significant relationship between methane and distance to conventional oil/gas wells?

Methane Concentration vs. Distance to nearest already-drilled-gas wells, Unconventional (left) vs. Conventional (right)

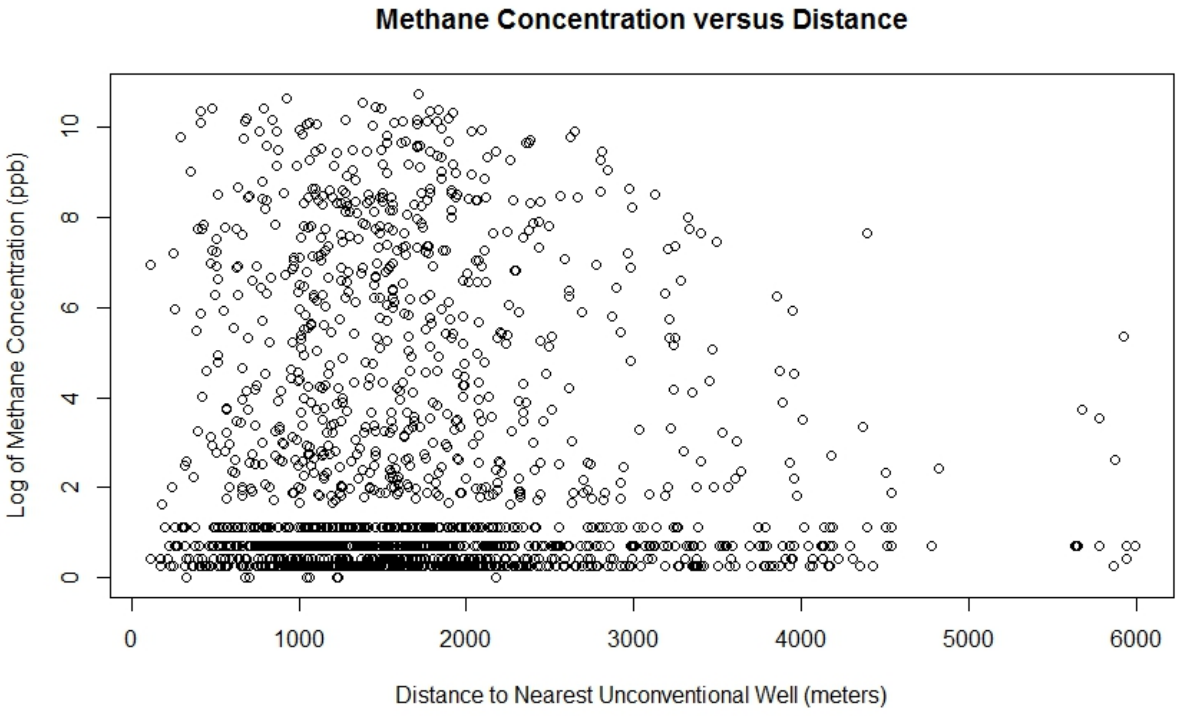




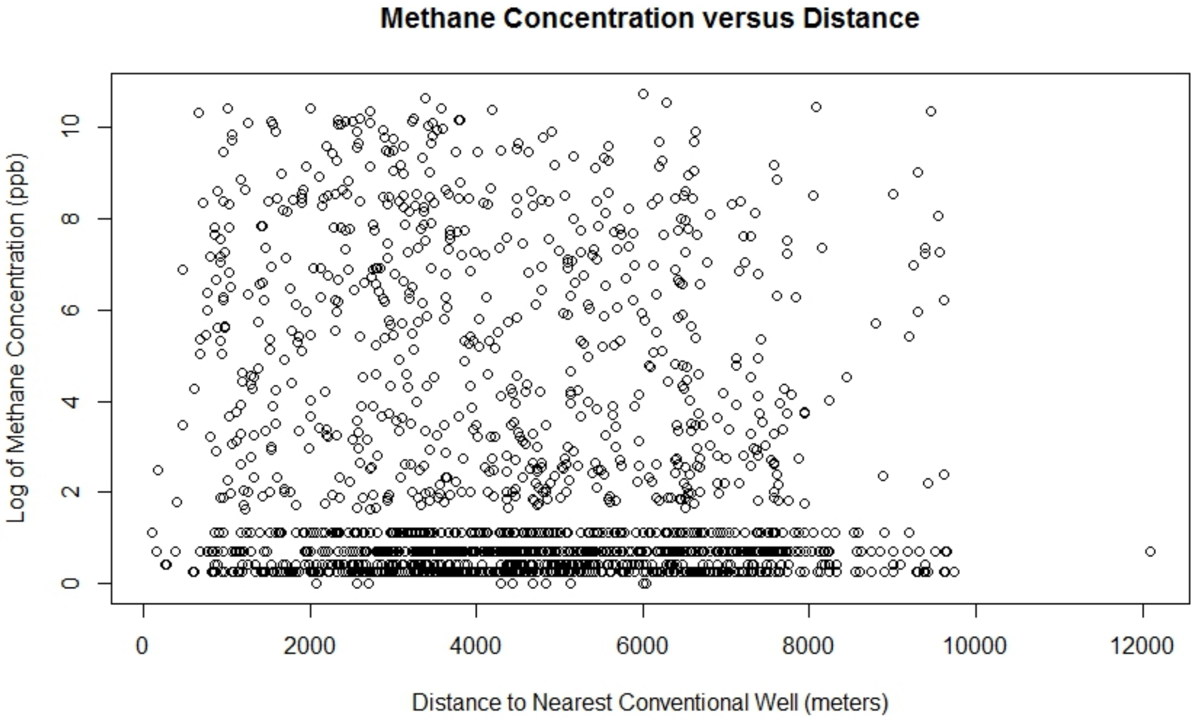
Nearest Already Drilled Unconventional Gas Well

Nearest Already Drilled Conventional Gas Well

Methane Concentration vs. Distance to closest-already-drilled-gas well (bootstrapped)



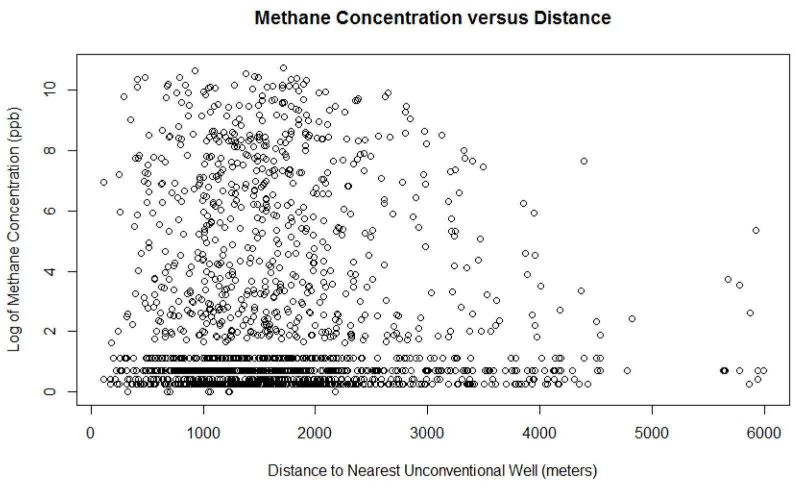
Closest Already Drilled Unconventional Gas Well

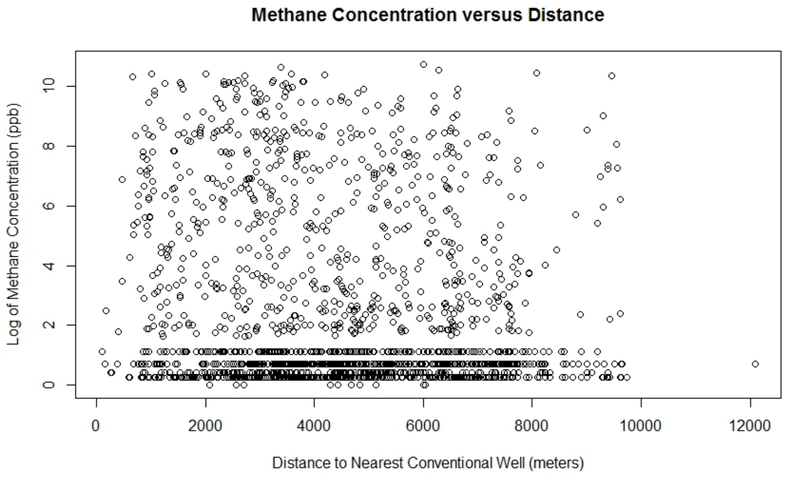


Closest Already Drilled Conventional Gas Well

Methane Concentration vs. Distance to closets already-drilled-gas wells (bootstrapped): Statistical Correlation

	Unconventional	Conventional
Pearson's Correlation	-0.061* p-value:0.0115	-0.097*** P: 5.8e-05
Spearman's Correlation	-0.018 P-value: 0.4452	-0.098*** P-value: 4.8e-05
Kendall's rank correlation	-0.012 P-value: 0.443	-0.067*** P-value: 7.38e-05
Regression analysis	-0.2968* P-value: 0.0115	-0.2085*** P-value: 5.865e-05





Conclusion:

1. Methane concentration in water is higher closer to gas wells (either unconventional or conventional): we see a significant but weak negative correlation.
2. The correlation is stronger in unconventional wells compared to conventional wells…but dataset has very very few conventional wells