An Analysis of Spatial and Temporal Methane Anomalies in Arkansas, Louisiana, and East Texas

Michael Feron [*1], Mark Holswade [†1], and Harrison Magee [‡1]

[1]Colorado School of Mines
Department of Applied Mathematics and Statistics

May 4, 2020

Abstract

The purpose of this paper is to provide analysis on the incidence and spatiotemporal distribution of methane levels in the Haynesville Shale region of East Texas, Arkansas, and Louisiana. Using data gathered from NASA's TROPOspheric Monitoring Instrument (TROPOMI), we analyze spatial and temporal trends in methane concentration in an effort to construct an accurate description of the natural and artificial prevalence, production, and evolution of methane in this region over time. In addition, we place particular emphasis on identifying locations of anomalously high, possibly artificial, methane production and on the correlation of methane in these regions with shale gas production. We address issues concerning bias in the TROPOMI data and perform careful analysis of the quality of the data. We make use of time-series and spectral analysis, and we attempt to account for the role that wind plays in concealing areas of methane production. Lastly, a regression model is constructed which correlates shale gas production with observed methane levels in the region surrounding the Haynesville Shale.

[*]mjferon@mymail.mines.edu
[†]mholswade@mymail.mines.edu
[‡]hmagee@mymail.mines.edu

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background Information

Methane, among other gasses, is one of the primary contributors to the greenhouse effect [6]. While greenhouse gasses are a necessary component of the Earth's atmosphere, their overabundance is currently serving to accelerate global warming. Methane is of particular interest because of its well-documented capacity to trap radiation in the atmosphere in an efficient manner. In fact, according to the EPA, methane has a global warming potential (GWP) – a measure of a gas's ability to absorb energy, measured relative to CO_2 – of 28-36 over 100 years [6]. For this reason, much current research has been devoted to the effects of methane emissions, as well as to their mitigation [4] [5].

In addition to the universal consequences of a warmer Earth, the release of methane is of practical interest to companies which transport and produce it. In general, lost or wasted methane results in lost revenue.

An eventual solution to this global problem will require, in particular, that any potential sources of greenhouse gas emissions be investigated. The primary focus of this paper is on the concentration and distribution of methane in the Haynesville region of Arkansas, Louisiana, and eastern Texas. This geographic region contains the Haynesville Shale, one of the most prolific sources of natural gas – whose primary component is methane – in the United States, as shown in Figure 1. The Haynesville Shale is depicted in navy blue, and its current shale gas production comprises about 14% of the total production in the United States.

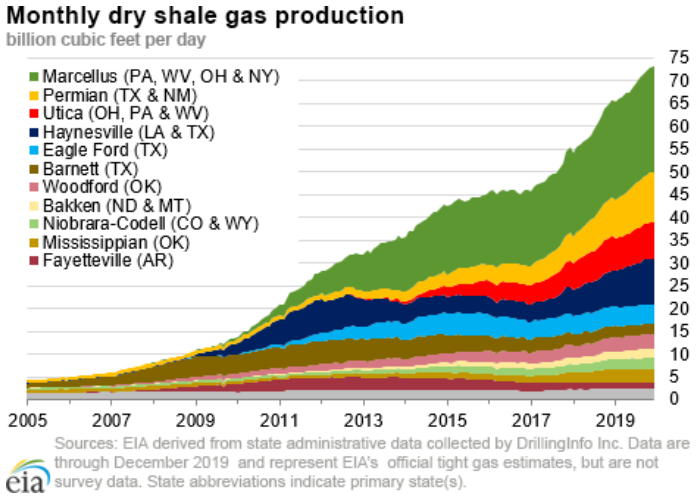
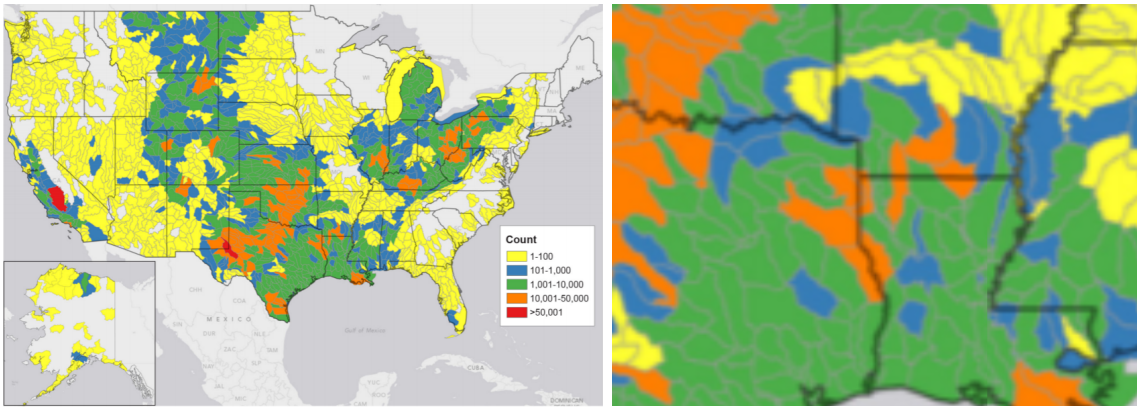


Figure 1: Top Shale Gas Producers [14].

One way to obtain methane as an energy source is through the extraction of shale gas. Shale gas is the result of organic matter from the ocean floor becoming trapped under sedimentary shale rocks and decomposing into methane and other gases [10]. Shale gas production in the United States has increased twelve-fold in the last decade, and this trend is expected to continue into 2035, resulting in over half of all U.S. based gas being sourced from shale. Shale gas formations are considered "unconventional" reservoirs due to their high difficulty of extraction. Typically, fracking methods are used to disrupt the shale formation and allow gas to escape [10]. This gas, however, can escape

unintentionally and release methane into the environment.



(a) United States (b) Region of Interest

Figure 2: Number of Oil and Gas Wells by County in the U.S.[1]

1.2 Research Question

In particular, we concern ourselves with the spatial distribution of methane in the Haynesville region and its evolution over time. Additionally, in an effort to pinpoint locations of unusually high methane output, we identify several spatiotemporal anomalies in methane concentration. Methane production can be linked to many possible sources, but our report is focused on methane production from highly populated metropolitan areas, wetlands areas, and methane gas release from natural gas extraction. Our paper aims to answer the question, "Which of these factors contributes to methane production, and how does this contribution vary through space and time?"

2 TROPOMI Data

The methane data in this study is collected by TROPOMI, an instrument developed by the Netherlands and the European Space Agency (ESA) on the Copernicus Sentinel-5 Satellite (S-5). S-5 P is "a single-payload satellite in a low Earth orbit that provides daily global information on concentrations of trace gases and aerosols important for air quality, climate forcing, and the ozone layer." TROPOMI is a spectrometer with spectral bands in the UV, visible, near IR, and IR ranges, allowing for the detection of various molecules, including methane [15]. TROPOMI's selected wavelength range, as well as its pairing with GOME-2 (another instrument), allow for diurnal variations which are ultimately translated into many gaseous readings, including methane.

2.1 Data Collection

We retrieved data from the dates between May 1, 2018 and March 31, 2020 within the polar rectangular spatial subset enclosed by these corner coordinates: -98.679°E, 29.498°N, -88.792°E, 35.518°N. This region is depicted in Figure 3, with the Haynesville shale region overlaid in light brown.

[1]Over 50% of all oil and gas wells in the U.S. have used fracking [10].

As can be seen in Figure 3, this region encompasses a far larger area than just the Haynesville shale. We choose a larger region to establish a background in this area upon which to construct our anomaly detection techniques. By comparing local instances of high methane concentration to an established background level, we can more accurately connect these anomalies to their geographic sources. Figure 3 shows the geographic region where shale gas extraction occurs in and around the Haynesville region.

Figure 3: The Haynesville-Bossier Shale[2]

2.2 Data Cleaning and Methodology

In our initial exploration of the data, we made a few observations about its contents. One piece of documentation recommends any data with a quality rating (`qa` value) less than 0.5 be ignored [2]. Analysis of `qa` values will include various scaling of the models using data that is deemed "acceptable" all the way down to data with `qa` values above 0. We also must take into consideration `qa` values are discrete, meaning there are only a few categories that correspond to certain weather events and satellite capabilities, not a continuous stream of quality based on some criterion function.

The second observation was the presence of "filler values", as described in the `netCDF` metadata, which appeared in approximately 90% of our methane mixing ratio observations. It should be noted that once our data is stripped of these invalid measurements, it contains solely observations whose `qa` values are either 0.4 or 1.0.

Finally, every methane measurement consists of two values: a raw measurement and a bias-corrected one. Henceforth, any reference to or computation involving methane measurements will use exclu-

[2]Map generated with `ggmap` [8]. Shapefile data courtesy of the U.S. Energy Information Administration [13].

sively the bias-corrected values. These values are corrected for various measurement errors that are outside the scope of this report.

3 Initial Analysis

3.1 Spatiotemporal Analysis

3.1.1 Four Quadrants

Here, we aggregate our data – by day and by week – in order to better understand how the presence of methane in the Haynesville region varies with respect to time and location.

We first divide our region into four spatial quadrants of equal size, approximately 154,706.2 km^2. These quadrants are oriented in row-major order; i.e. `Q1` is the northwest quadrant, `Q2` the northeast, `Q3` the southwest, and `Q4` the southeast. Any further division of data discussed in this paper will adhere to this convention. Figure 4 shows the Haynesville shale with grid lines demarcating each quadrant.

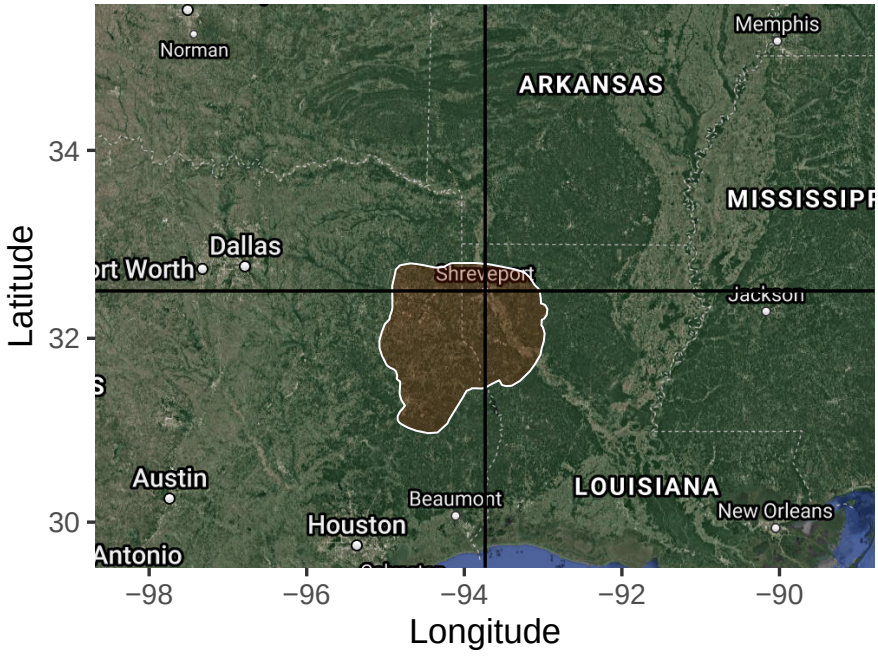


Figure 4: Four quadrants and the Haynesville Shale

Figure 5 and Figure 6 show plots of bias corrected methane mixing ratio vs. time between May 1, 2018 and March 31, 2020 – computed daily and weekly, respectively, for each quadrant.

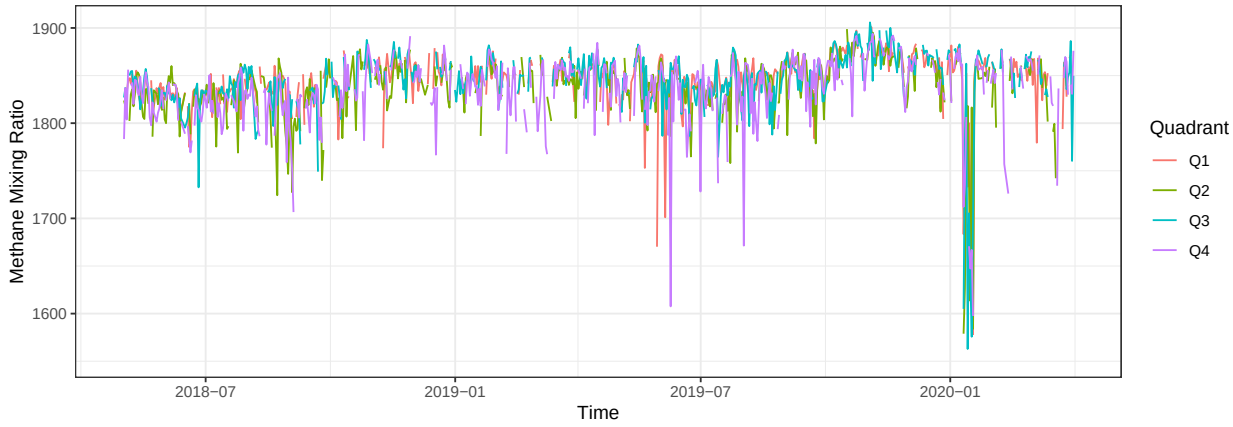


Figure 5: Mean daily methane mixing ratio, by quadrant

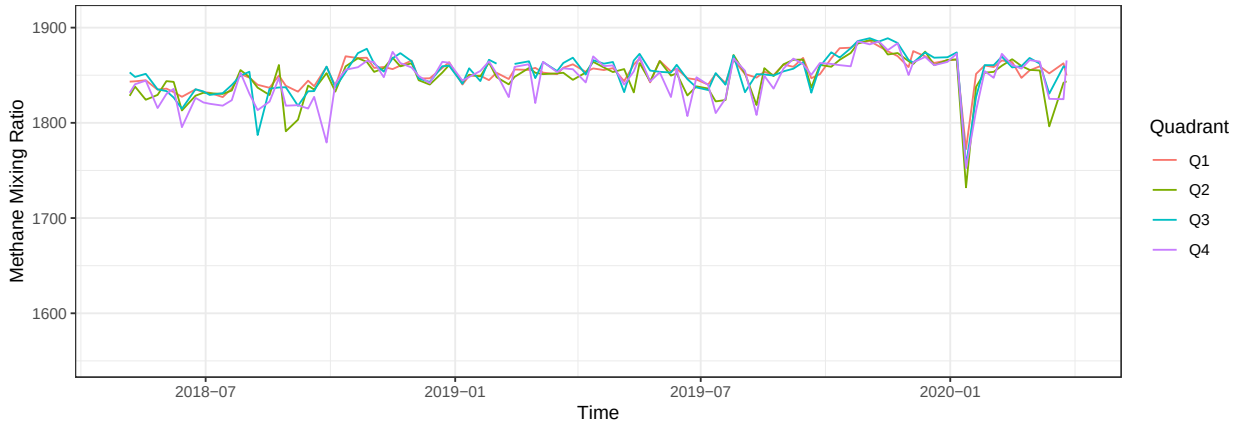


Figure 6: Mean weekly methane mixing ratio, by quadrant

The most glaring feature of this plot is the tremendous drop in methane mixing ratio in the second week of January of 2020. Our research indicates no events in the United States or otherwise that would cause methane levels to drop at such a rapid rate. That each quadrant appears to drop in tandem could indicate mechanical malfunctions, but the documentation states nothing of the sort. The `qa` values during this time are almost exclusively 0.4, which provides evidence that there was likely some sort of instrumentation failure. With the information available, this strange behavior in the data is rather inexplicable, and so we exclude such data from further analysis.

The next things to note are the conspicuous gaps in Figure 5. These are missing values, an indication that on some days, and in some quadrants, valid observations were not collected. Figure 6 contains only one such break. Such breaks may be seen as a result of persistent cloud coverage over a region or other issues concerning the collection of valid methane measurements. This is obviously problematic when trying to do daily analysis – these data do not exist; however, there is also a more subtle problem that lies with weekly analysis. The complete absence of daily data in some areas and at some times reflects quite clearly the sparsity of data, especially at high spatiotemporal resolutions. This presents a problem by potentially introducing bias and high variance in some of our weekly measurements, something we are careful to avoid as much as possible in subsequent analysis.

Furthermore, in Figure 6, there is a notable homogeneity in methane ratios with respect to each quadrant. At first glance, there is no single quadrant which deviates substantially from the overall trend in time. This is very likely a result of our imposed spatial resolution – four quadrants – being too coarse. Any spatial anomalies which might be present are being washed out by the background noise within their quadrant. Later, we tune our grid to a finer resolution to account for this.

With respect to an overall pattern in time, Figure 6 offers a visual indication of a few potential anomalies. First, it should be noted that, on average, methane levels are higher from mid-October through mid-November. In addition, there are several dips in methane mixing ratio over the course of the 23 months of collected data, particularly in quadrants two, three, and four.

Table 1 depicts summary statistics – sample size, mean, standard deviation, inner-quartile range, and mean absolute deviation – of methane mixing ratio by quadrant, taken over all observations in time.

Quadrant	n	μ	σ	IQR	MAD
1	275,794	1856.6	27.0	25.4	18.8
2	227,305	1857.0	33.1	26.1	19.4
3	242,513	1859.2	35.2	26.6	19.6
4	183,182	1858.6	36.2	27.1	20.0

Table 1: Four-Quadrant methane summary

We see that `Q3` and `Q4` have the highest mean methane mixing ratio values. These quadrants roughly correspond to the part of our region containing Louisiana, southern Mississippi, and the portion of Texas lying south of Dallas-Fort Worth – this region includes most of the Haynesville shale. More than likely, the differences in mean highlighted here are not statistically significant, and with such large regions considered, it is not surprising that the differences here are negligible.

Now, this analysis does not account for seasonal trends, or even monthly or daily fluctuations in methane level. We'll address these aspects later in the paper.

3.2 Proportionality Significance to the Mean

To better account for fluctuations in time, it is useful to compare methane concentrations in each quadrant on a month-by-month basis. To do this, we examine the methane mixing ratio in each quadrant every month, and we compare its value to the mean value over the entire region – a monthly global baseline mean. We then compute an estimate for the true proportion p, the statistic $\hat{p}$, for the proportion of months for which each quadrant demonstrates above-average methane levels with respect to the global baseline. Table 2 contains a summary of these results.

`Q1`:	0.39	`Q2`:	0.30
`Q3`:	0.78	`Q4`:	0.35

Table 2: Monthly Elevated Methane Levels

These results paint a somewhat different picture than those given in Table 1. Although `Q3` is shown to have the highest mean methane concentration as well as the largest proportion of months with

above-average methane levels, this same consistency is not present in the other quadrants. `Q4`, which has the second-highest mean methane concentration, spent only 8 of the observed 23 months with a mean methane level above the global baseline. This is an interesting pattern, perhaps indicating that the levels of methane were abnormally high in those eight months where `Q4` did demonstrate above-average values. Conversely, this may indicate that much lower measurements were recorded in `Q4` during the other 15 months. This would seem more likely, at least upon visual inspection of Figure 6. This is also supported by the summary statistics provided in Table 1: `Q4` has less measurements over time, and a higher standard deviation, IQR, and MAD than any of the other quadrants, all indications of higher variation in this quadrant.

Ultimately, however, the key takeaway here is that although each quadrant demonstrates very similar mean methane levels over time, there is enough variation present in the data to yield different results when considering the *proportion* of time spent above a time-localized mean.

4 High Spatial Resolution

We now impose a much finer spatial resolution of 10,000 sectors on our region. Each sector is approximately 61.9 km^2 in size where the shape is rectangular with respect to latitude and longitude. We examine methane mixing ratios in each sector on a *daily* basis. We again compute an estimate for the proportion of days for which each sector demonstrates above-average methane levels with respect to a daily global baseline.

The results are more conveniently examined visually. Figure 7 depicts the spatial grid structure, while Figure 8 contains a heatmap whose values indicate this estimated proportion for each sector.

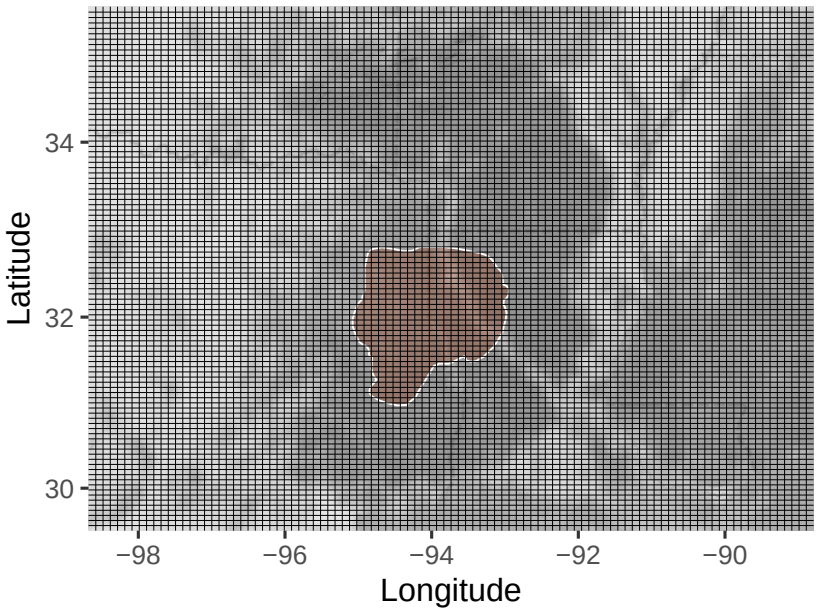


Figure 7: 10,000 sectors and the Haynesville Shale

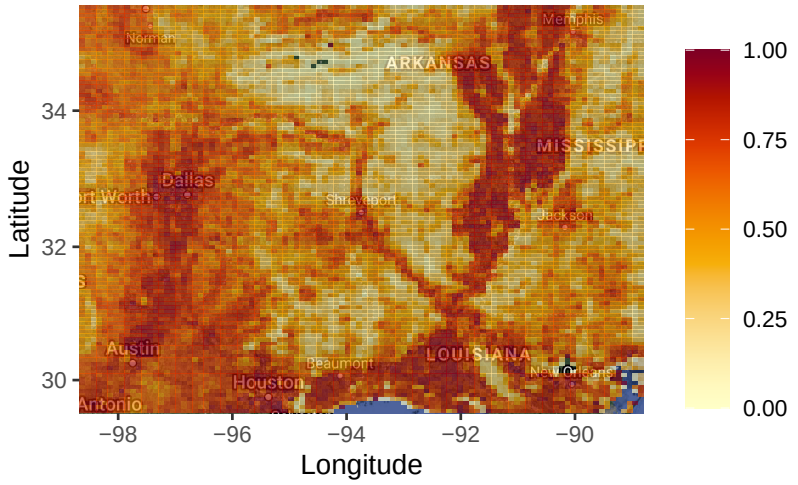


Figure 8: Estimated proportion of days with above-average CH_4 levels

This figure can be thought of as depicting those sectors which demonstrate time-persistent, above-average methane concentrations, affording us a sense of the overall spatial composition of methane in the region.

Now, when compared to the satellite view of our region in Figure 3, there is a stark and uniform pattern to be gleaned. Almost unfailingly, regions with large estimates for $\hat{p}$ correspond to the 'light-green' regions in Figure 3, while low estimates for $\hat{p}$ correspond to 'dark-green' regions. This pattern persists across the whole of our observed geographic subset.

We see particularly high proportions in the areas near the Mississippi River, the wetlands in southern Louisiana, and the regions connecting Dallas, Austin, and Houston – all of which are major metropolitan areas. In addition, and of particular geographic interest, is the concentration of high proportions in the corridor in Louisiana running southeast from Shreveport to New Orleans. Also present is a distinct set of high values in the Haynesville shale region.

5 Data Masking

When evaluating methane data, a major issue with the TROPOMI instrument is its inability to consistently and correctly handle data above water [15]. With this in mind, we see major areas of water in our region which include the Mississippi River, the Mississippi River Delta, the Red River, and the Gulf of Mexico, as well as some other minor lakes and tributaries. In our data, we see a small, but not insignificant, amount of data above water that could possibly introduce bias into our measurements. In many cases, it is apparent that the provided `qa` values do not sufficiently account for the presence of water. Furthermore, considering the high proportion of days with elevated methane in many of the wetlands regions of Louisiana and Mississippi – as examined in Section 4 – the potential existence of bias in our data must be more thoroughly explored. One method to account for this bias is the application of a land-water mask.

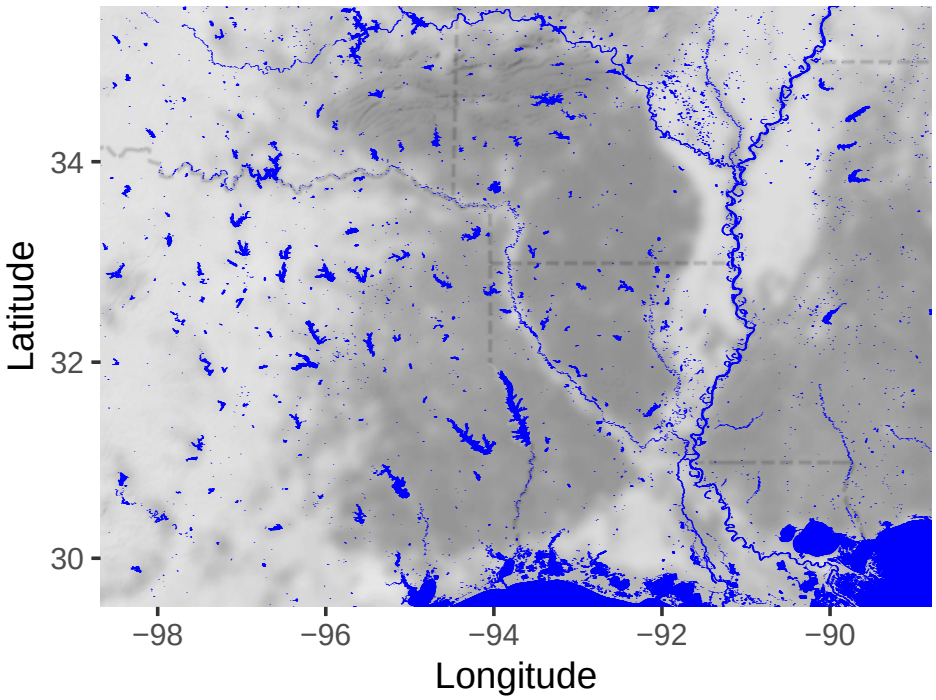


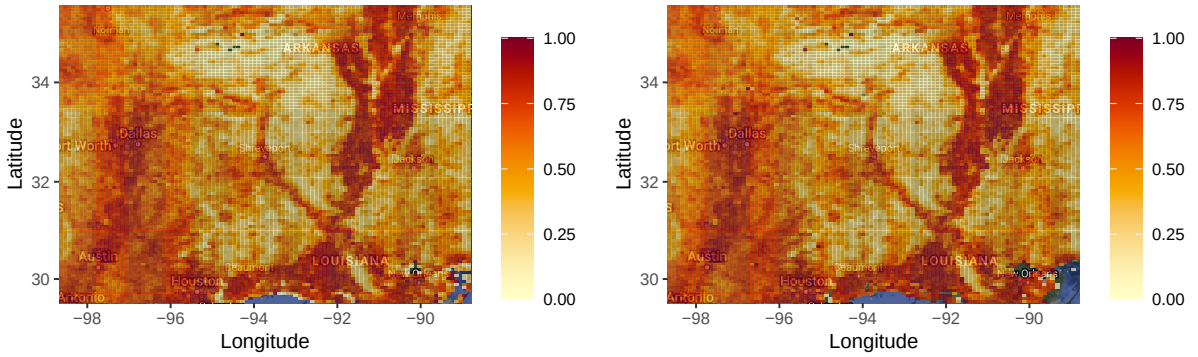
Figure 9: NASA MODIS Land Water Mask

5.1 Methodology and Considerations

By masking the data points, we classify each data point as either above water or above land, and we discard the former. We then re-perform analysis based on the corrected data. We use publicly available data obtained from NASA's MODIS MOD44W v006 Land Water Mask to accomplish this [9]. Figure 9 depicts this mask overlaid on our region.

This data provides readings on an irregular – or non-rectangular – spatial grid with 250 meter resolution, with binary values indicating the presence of water or the lack thereof. To apply this mask, we fit a custom spatial grid to the mask data containing 2875×2875 sectors. At this particular resolution, the grid fits relatively well with respect to the structure of the provided mask: each sector contains at least one mask point, with the majority (99.7%) containing exactly one or two mask points, and only trace percentages of sectors containing either three or four points. Data within sectors containing any points with indicators for water are flagged for removal.

With the water mask applied, the analysis from section 4 is performed again. A new heat map representing the proportion of days for which each sector demonstrates above average methane levels with respect to the daily baseline is presented in Figure 10.



(a) Without mask

(b) With mask

Figure 10: Land water mask: estimated proportion of days with above-average CH_4 levels

This area undergoes almost no change with respect to the proportions of days above the mean when the mask is applied to it.

5.2 Analysis of Mask

A visual inspection of the two plots suggests that applying the water mask caused almost no change in the analysis, nor in the overall features of the heat map. This is consistent with the size of the mask, since the mask only removes approximately 1.16% of the data used by the previously constructed heat map. Removing such a small portion of the overall data set is unlikely to significantly change the results of the analysis. All further analysis will be performed only on the masked data set. These results should still be relevant to the unmasked data, as the two data sets are so similar.

It is important to consider that this analysis likely did not eliminate bias from all measurements, particularly those made near wetlands regions. The water mask data used above, being implemented at 250 meter resolution, is likely too coarse to account for smaller areas which might contain wetlands.

6 Quality Flag Analysis

6.1 Spatial Distribution

In addition to bias which may be present as a result of data collected on or near bodies of water, it is important to consider the effect that utilizing data with low `qa` flags might have on our analysis. As discussed previously (Section 2.2), our data consists of measurements with `qa` flags of either 1.0 or 0.4. The former rating indicates data measured with high confidence, and the latter measured with lower confidence. We refer to these low confidence measurements as marginal ones. Approximately 37% of our data, with the land-water mask in place, are associated with a high `qa` value. The remainder of the data, about 63%, consist of marginal measurements.

Up to this point, our analysis has treated all data without consideration for their associated quality flags. However, given the possible existence of bias in some measurements, and given that the majority of the data are marginal, it is necessary to consider the effect that removal of marginal measurements might have on any inferences drawn about the data.

Figure 11 is an image plot depicting the spatial distribution of the proportion of total observations with high quality flags.

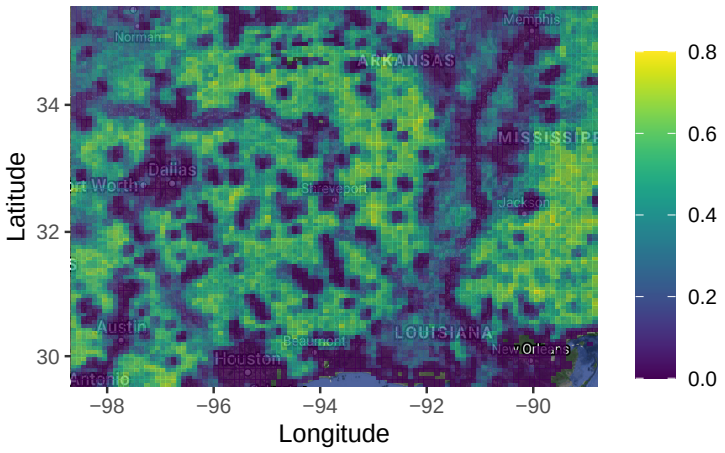
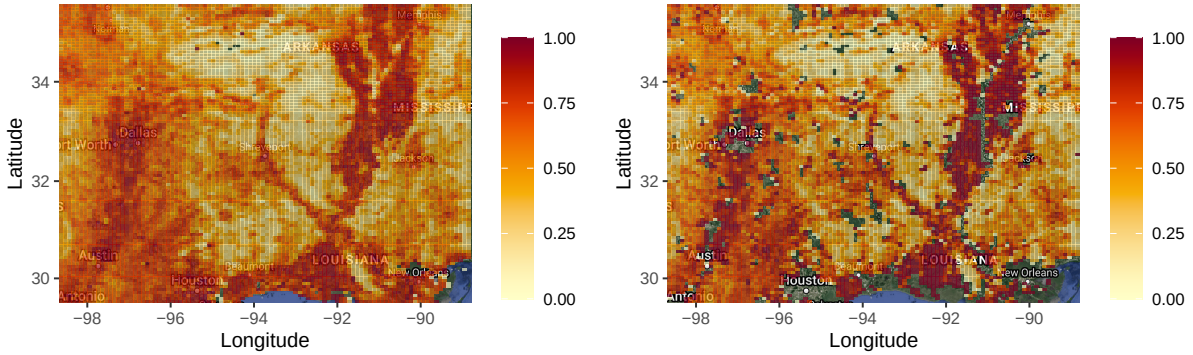


Figure 11: Proportion of observations with high `qa` flags

We see, particularly in the wetlands regions and areas surrounding the Mississippi River, a correspondence between low quality flags and high estimates for the proportion of days with above-average methane levels (as presented in Section 4 and Figure 8).

6.2 Proportion Analysis

As suspected, the majority of low quality measurements were made in areas where the presence of water poses especially high risk to the integrity of the data. For this reason, we now re-perform the analysis of Section 4 – using only high-quality data, with the mask applied – computing an estimate for the proportion of days for which each sector demonstrates above average methane levels with respect to the daily baseline. A comparison of the results is shown in Figure 12. The figure on the right depicts the updated results using only high-quality data. The figure on the left depicts the prior results. Both analyses use only data with the land water mask applied.



(a) Marginal `qa` included

(b) High `qa` only

Figure 12: `qa` analysis: Estimated proportion of days with above-average CH_4 levels

Again, the data is almost unchanged, aside from sparsity. These two figures are very similar in overall structure. The same general regions – the wetlands, the corridor running through Louisiana, etc. – are all highlighted here for demonstrating time-persistent methane levels. There appears to be an even starker contrast between areas with high proportion estimates and those with low estimates. It would be wise to be wary of this difference, however, as with such a substantial portion of the data removed, many of these estimates for $\hat{p}$ are made on the basis of very few observations. We see that this is the case in general, as many sectors contain no valid high-`qa` observations for the entirety of the 23-month period of observation.

Overall, however, the results obtained using exclusively high quality observations are quite consistent with those obtained when including the marginal data. As such, we conclude that levels of methane are consistently above average in the following regions: the metropolitan sprawl between Dallas, Austin, and Houston; the areas surrounding the Mississippi River and Mississippi River Delta; the areas surrounding the Red River; the corridor in Louisiana running southeast from Shreveport to New Orleans; and the wetlands regions of southern Louisiana.

7 Wind Analysis

7.1 Motivation and Method

The effects of wind moving methane acts as a confounder for identifying regions of methane production. It is difficult to separate what methane is produced in a region and what is being blown in from another region. It is probable that many of the high methane areas are simply areas where methane is being blown in from other regions.

In order to separate these effects, a simple method for removing the effect of wind movements was devised. Using the wind speed and direction columns given in the TROPOMI data set, we create a new data set by moving each data point by its wind vector over a 24 hour period. To illustrate the concept, if a data point from the TROPOMI data set was read on January 1st at 12:00am, with a latitude/longitude of (60,60) and a wind vector in latitude/longitudinal units of (1,1), the data point would be shifted to January 2nd at 12:00am with a latitude/longitude of (84,84).

Both the original and wind shifted data sets were then split into a 50×50 grid and averaged on methane mixing ratio by day. The wind adjusted methane mixing ratio for each grid point and day was subtracted from the corresponding grid point and in for the original data set. Grid point and days that did not have data for both data sets were excluded. Altogether, approximately 77% of data was excluded from the final data set. This new data set was averaged on methane mixing ratio over all time and plotted.

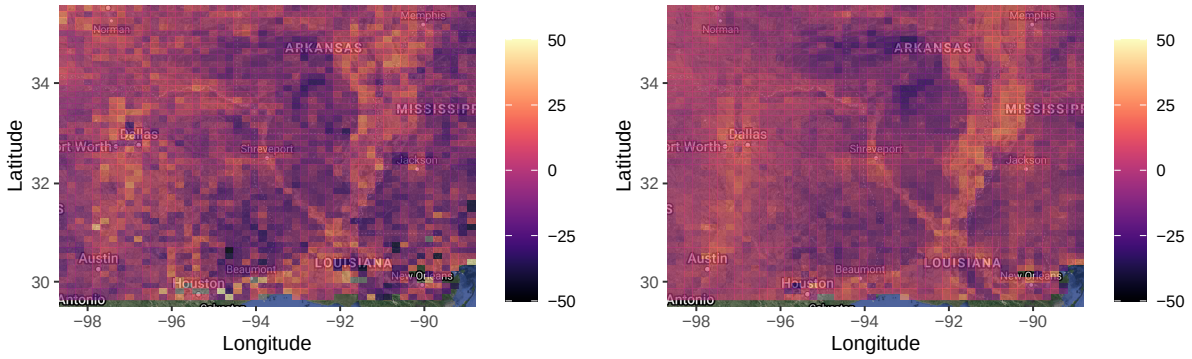
The hope is that by subtracting away the methane concentrations and movements from the previous day, the methane production in each grid point over a 24 hour period can be isolated. Although the method is very simple, we believe that it can give a rough picture of the average methane production in the region.

This method resulted in significant data loss. 34% of the data points were shifted outside of our region in the wind adjusted data and therefore removed from the analysis. In addition, two full days of data were removed from the front and back of the final data set, resulting from the lack of

overlap between the original and wind adjusted data. There was also a large portion of data loss caused by lack of overlap in quadrants on any given day, resulting in a loss of 77% of the remaining data. Altogether, this method was only able to make use of 15% of the total data set available. Much of this data loss is intrinsic to the method and therefore unavoidable, but some data loss was mitigated by reducing the resolution of the produced quilt plot to 50×50, as compared to the 100×100 quilt plots generated in other sections. By using a lower resolution, more quadrant/day combinations were filled in with data, resulting in more of the data being preserved in the final data set. Each quadrant was calculated with an average of 52 data points.

7.2 Results of Analysis

A plot generated using the method described in Section 7.1 is presented in Figure 13a, with a plot of average methane concentrations over all time centered around the global (regional) mean presented in Figure 13b for comparison.



(a) Estimated average methane production

(b) Average methane concentrations

Figure 13: Comparison of methane production vs concentrations over all time

Figure 13a shows a picture of methane production very similar to that of methane concentrations shown in Figure 13b. It also clearly identifies the regions identified in Figure 8 as being areas with high proportions of above average methane concentrations as also being heavy methane-producing areas. The wetlands region and Dallas-Fort Worth to Austin corridor identified in Section 12.2 stand out as the most obvious regions of methane production. The plot also indicates a large amount of methane produced following the Red River, which spans from the Texas-Oklahoma border down to Shreveport and connects with the Gulf of Mexico. These results seem to indicate that methane is heavily produced in very wet areas, where a lot of plants are able to grow and eventually die and decompose, as well as areas of heavy urban traffic, including Interstate 35 (I-35) between Austin and Dallas-Fort Worth. The Haynesville Shale region also stands out in the map, being highlighted as an area of moderate methane production. Justification for the validity of the existence of methane "sink" areas, such as the area between the Red River and the wetlands region, has not been researched.

7.3 Conclusions and Potential Improvements to Method

It is important to note that all justification for the effectiveness of this method is based off available information and previous analysis performed. The results have not been compared and confirmed

with outside data, and further confirmation is required to prove the method's effectiveness.

This analysis produced mixed results. We hoped this model would be able to isolate some areas of methane production that were not immediately obvious from Figure 13b. This does not appear to be the case. Figure 13a does show a starker contrast between the wetlands region and Red River as well as between the Red River and the Dallas-Fort Worth to Austin corridor when compared to figure 13b. The validity of this is difficult to assess without outside data to compare it to, but this may be due to diffusion of methane in the atmosphere, assuming the model is effective.

The next step for the method would be to find an outside source of data to justify its effectiveness. If the method was deemed effective, a possible improvement could be using outside wind data on a finer time scale to move the data points in a more realistic way. In theory, the actual spot the to which the methane moved could be accurately determined by moving it based on multiple wind vectors, e.g. an hourly basis. Another possible refinement would be more thoroughly considering the effect of methane being blown into a single region from multiple locations. The method used simply averages any data points that end up in a region, and does not consider the possibility that regions with methane being blown in from many locations would have a higher concentration than areas with methane blown in from only a few locations.

8 Time Series Considerations

Time is one seemingly very important variable, responsible for presumably quite a bit of the variance in our spatiotemporal data. Naturally, we conducted thorough time series analysis; however, this analysis lent no useful results. The explanation of analysis can be found in the Appendix, Section 12.2.

9 Haynesville Analysis and Regression

The Dallas-Fort Worth to Austin region and the Wetlands region have simple possible explanations for their high methane production: traffic and microorganisms. Dallas-Fort Worth to Austin runs along I-35, one of the most-traveled interstates in our region. Vehicular travel contributes massively to methane levels [7]. The likely Wetlands region contributors are mainly microorganisms along the Mississippi River, which also contribute substantially [1]. In the Haynesville region however, we seek to explain the high levels and behavior with a more interesting explanation: natural gas production, as posed in our research question in Section 1.2.

As previously discussed in Section 7, we compound our loss of data at each step in our wind analysis, so we use a new method.

9.1 Method

We began by isolating data that lies inside the Haynesville shale region, depicted in Figure 3. We took the daily mean in the area outside the Haynesville shale region and subtracted it from the each value in the Haynesville shale on that particular day. This functions as a background correction by removing potential sources of background noise and variation which do not originate from within the Haynesville shale. Figure 14 contains a comparison between the original data and the background-corrected data, in which each plot depicts the mean methane mixing ratio in

the Haynesville shale region: on the left is the original data, and on the right is the background-corrected data. The original data has been centered to produce a more useful comparison. Note that the mean values depicted here are taken over the entire period of time.

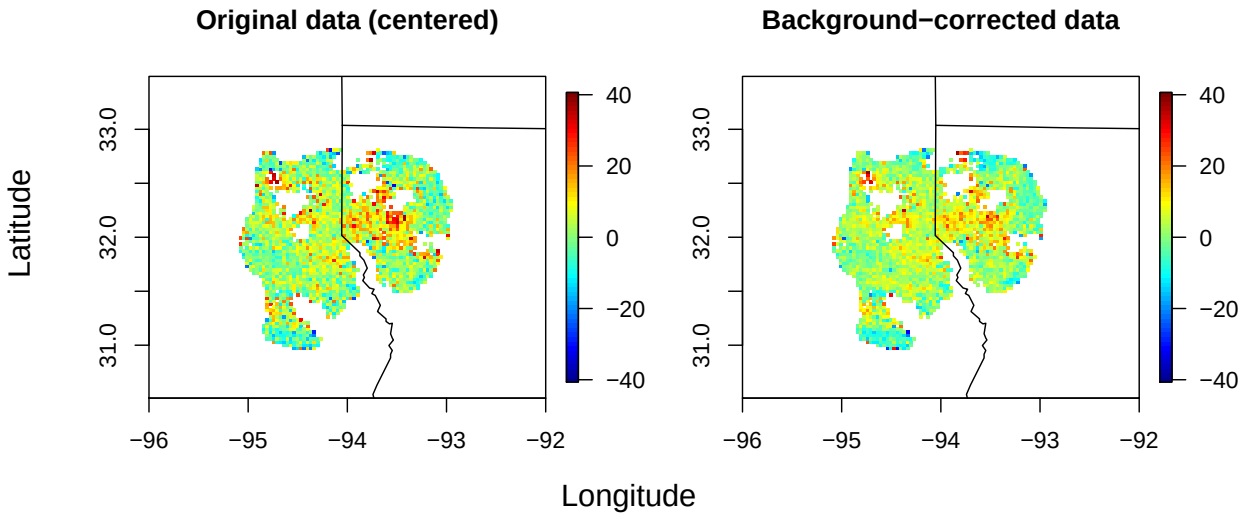


Figure 14: Mean methane concentration in the Haynesville Shale

We then aggregated this data by month and averaged it, thus leaving one background-corrected measurement per month inside the Haynesville shale region. Next, we imported a dataset from the U.S. Energy Information Administration on shale gas production in the Haynesville region. This monthly data ranges from May 2018 through February 2020, and includes an average daily shale gas production value, measured in billions of ft^3, in the region for each month [12]. We then fit a linear model relating methane mixing ratio to natural gas production in our region through time.

9.2 Results

The results are pictured in Figure 15,

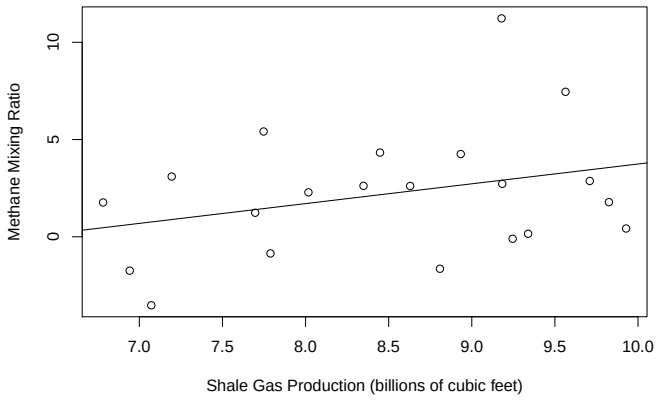


Figure 15: Linear Model for the Haynesville Region Background Corrected Monthly Values

and an output for the model summary is given in Table 3.

Coefficients:

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	−6.4345	6.1397	−1.048	0.308
shale	1.0176	0.7181	1.417	0.173

Residual standard error: 3.209 on 19 degrees of freedom
Multiple R-squared: 0.09559, Adjusted R-squared: 0.04799
F-statistic: 2.008 on 1 and 19 DF,

Table 3: Model Summary

We can see what appears to be a positive, linear relationship between production and emissions; however, our R^2 indicates we are performing very poorly. Figure 15 shows a positive linear relationship, but not one strong enough for us to be comfortable drawing conclusions - our Adjusted $R^2 = 0.048$, a very small value. Ultimately, we see a trend, but simply do not have enough data to support our suspicion that natural gas extraction contributes to methane leakage and production.

Since a trend appears to be present but the Adjusted R^2 is low, the next logical steps would be to collect or retrieve more data somehow. We could replicate each monthly measurement of shale gas four times to create pseudo-weekly measurements and aggregate the methane measurements by week, or we could linearly interpolate between the months to create unique weekly measurements in the shale gas measurements. Before doing this, we should evaluate the likelihood of success in doing so.

We first start by evaluating plots of our regressor and predictor. Figure 16 is a time-series plot of shale gas production in the Haynesville shale region, while Figure 17 is a time-series plot of monthly methane mixing ratios in the same region.

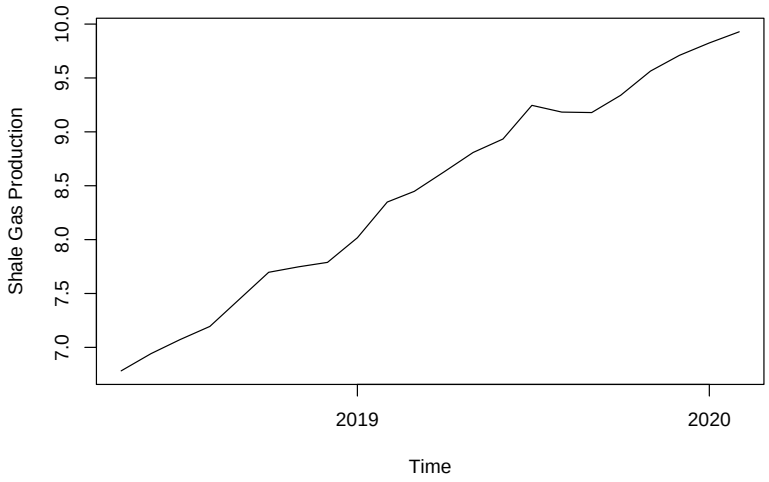


Figure 16: Shale gas production in the Haynesville region

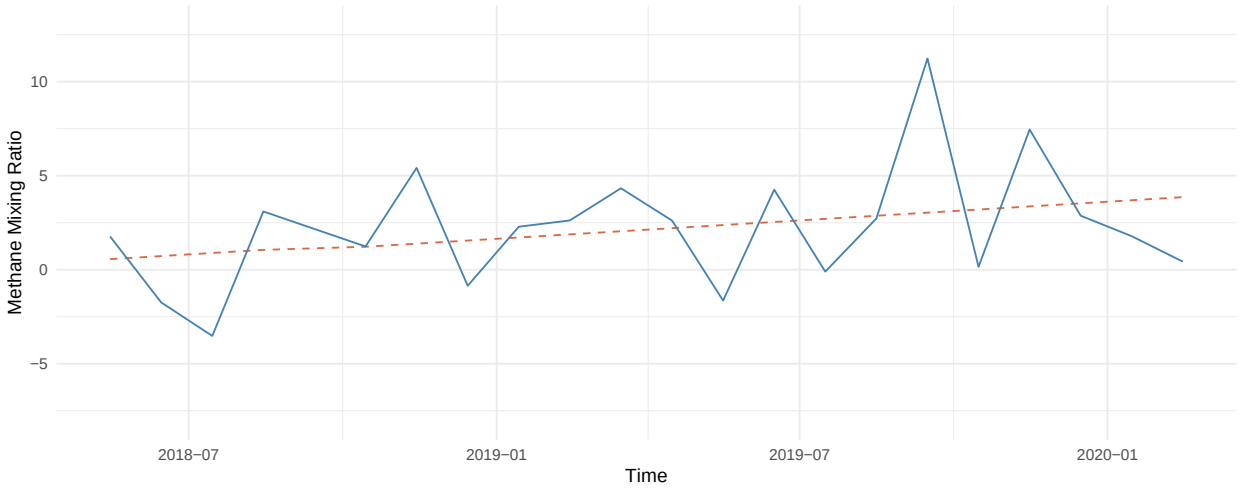


Figure 17: Haynesville methane mixing ratio

From these plots we see a very clear, nearly perfect linear trend in Figure 16, and a slight, positive linear trend in Figure 17. We know that much of the variance captured by the model is a confounding correlation with time, but we likely cannot improve this fit much without including time. Any bootstrapping method or artificial data acquisition from our shale gas data would not accurately capture fluctuations in our methane data without introducing unnecessary bias. We conclude that we cannot correlate shale gas production with methane mixing ratio alone.

While we might be able to produce a much better fit by introducing time, we feel this is not a necessary step to take. We are interested in correlating instances of high methane levels with associated instances of high shale gas production. While both variables increase over time, this does not provide us with useful inferences about our regressor and response. We ultimately wanted to directly correlate the two, but we see from the shape and activity in both, we likely could not do so accurately.

10 Spectral Analysis

Much of the time-series analysis performed up to this point has hinted at the existence of possible periodic trends in the data. Uncovering these trends can help to elucidate underlying correlation structures present in methane concentration over time. One way to accomplish this is through spectral analysis, which involves the decomposition of a signal into its constituent frequencies through the mapping of the signal into the frequency domain. In this domain, the analysis of periodicity becomes much more tractable.

Perhaps the most common method used in this type of analysis is the discrete Fourier transform, or the DFT. The DFT maps a discrete sequence of values, $\{x_n\}$, sampled from a signal, into a another sequence, $\{z_k\}$, which resides in the frequency domain. The transform is given by

$$z_k = \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n/N},$$

where N is the number of samples, and $n, k \in \{0, \ldots, N-1\}$. Each x_n is assumed to be sampled on equally spaced intervals of time.

To extract useful results from the Fourier transform, it is necessary to remove any general trends which might exist in a particular signal. Figure 18 contains two plots depicting mean daily methane concentrations, over time, over our entire region. The first plot shows the original signal, with its upward trend overlaid in red. The second plot shows the de-trended signal. Only data associated with high `qa` values is used here, as the marginal data introduce potentially problematic and unnecessary variation which could conceal the existence or relative strength of any periodic trends.

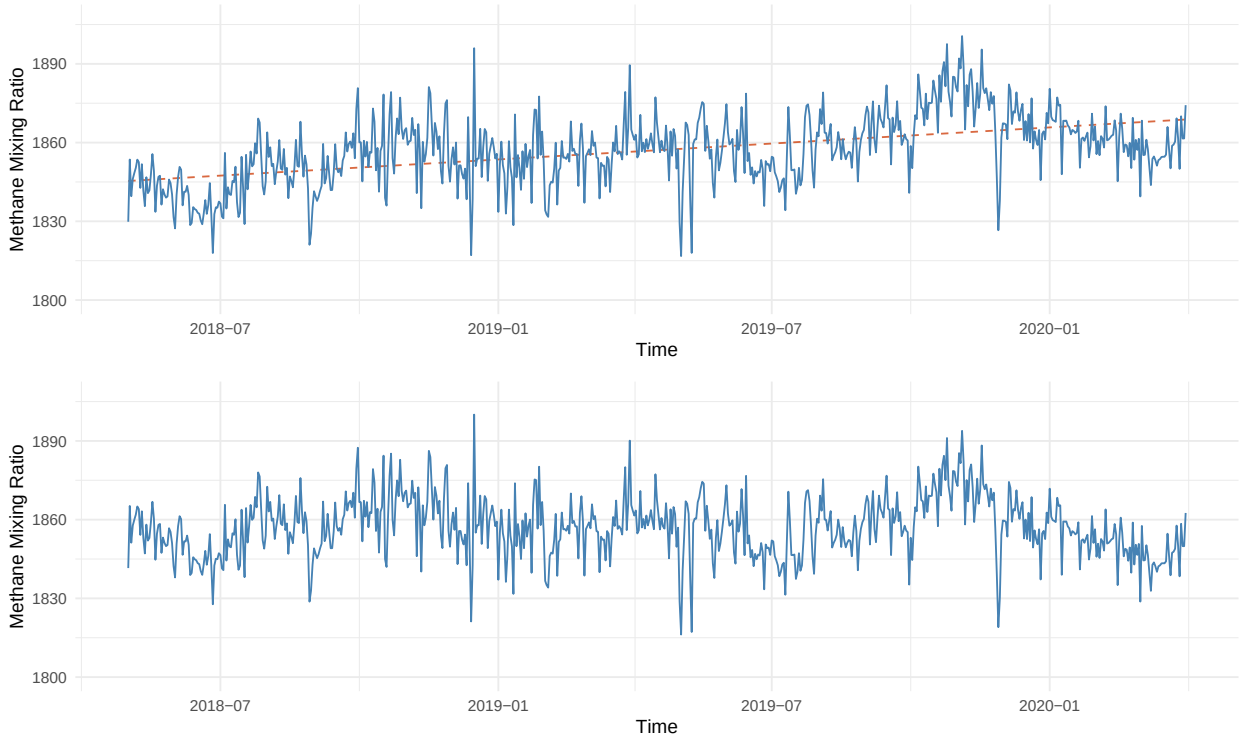


Figure 18: Mean daily methane mixing ratio

Methane concentrations on any days without valid observations are imputed using a centered moving average. The result, shown above, can be thought of as a sequence of 701 equally-spaced, daily samples from a signal, and thus as a candidate for the Fourier transform.

Figure 19 is a depiction of the frequency spectrum of the above methane signal, computed using the discrete Fourier transform. In it, the modulus of each z_k is plotted against its associated frequency, measured in months^{-1}. The modulus represents the strength of a signal at a particular frequency.

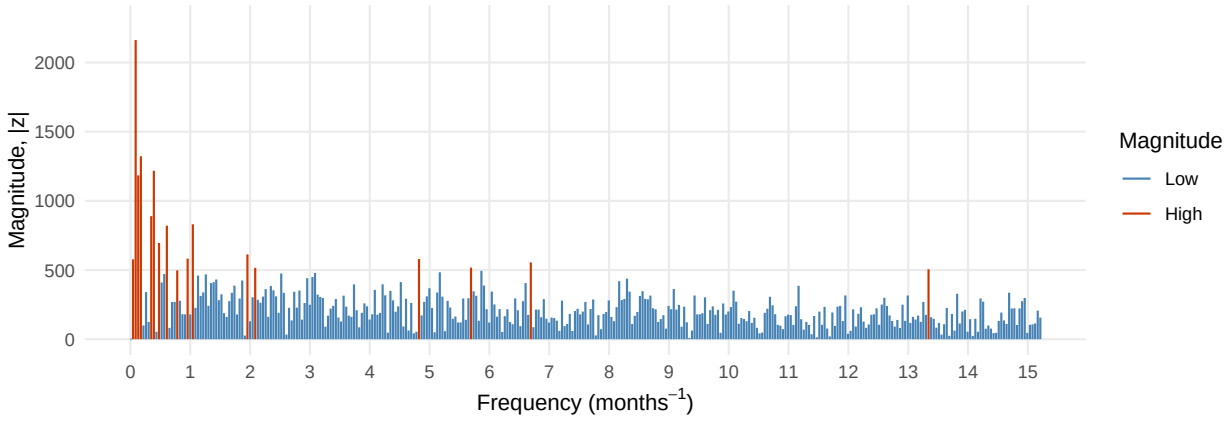


Figure 19: Frequency spectrum for daily methane mixing ratio over entire region

The top 5% of signals, in terms of strength, are marked in red. By far the strongest signals are those lowest in frequency. These frequencies represent relatively long-term trends in methane concentration over time. The lowest of these frequencies, at 0.04 months^{-1}, has a wave period of 23 months, and may represent a trend in methane which persists over multiple years. Unfortunately, since the collected data do not span an entire two-year period, it would be unwise to draw any firm conclusions about this specific trend. Moreover, other constituent frequencies are present with much greater magnitude.

As has been discussed previously, there is also an apparent seasonal component to methane concentration, with elevated levels observed in colder months and reduced levels observed in warmer months. The dominant signal in Figure 19 is that present at a frequency of 0.09 months^{-1}, with a corresponding wave period of 11.5 months. This signal is strongly indicative of this seasonal component, and more firmly supports our earlier speculation that the spring and summer seasons see a reduction in methane levels relative to fall and winter.

In addition to this annual trend, there are several notably strong sub-annual signals, namely those at frequencies of 0.13, 0.17, and 0.39 months^{-1} – with associated periods of 7.7, 5.8, and 2.6 months. These signals likely combine to create the seasonal variation we see in the data.

We now examine the frequency spectrum of methane signals confined to three areas of special interest: the Haynesville shale region, the wetlands, and the Dallas-Fort Worth/Austin metropolitan corridor.

10.1 The Haynesville Shale

When restricted to the small Haynesville region, the data are not sufficiently dense to warrant signal sampling on a daily basis. For this reason, weekly means are computed in lieu of daily means. Figure 20 contains time-series plots of trended and de-trended mean weekly methane concentrations in the Haynesville shale region using high-quality data.

Again, missing values are inferred using a centered moving average. It is crucial to note that, in this case, 21 of the 100 sampled weeks do not contain any valid high-quality observations. This

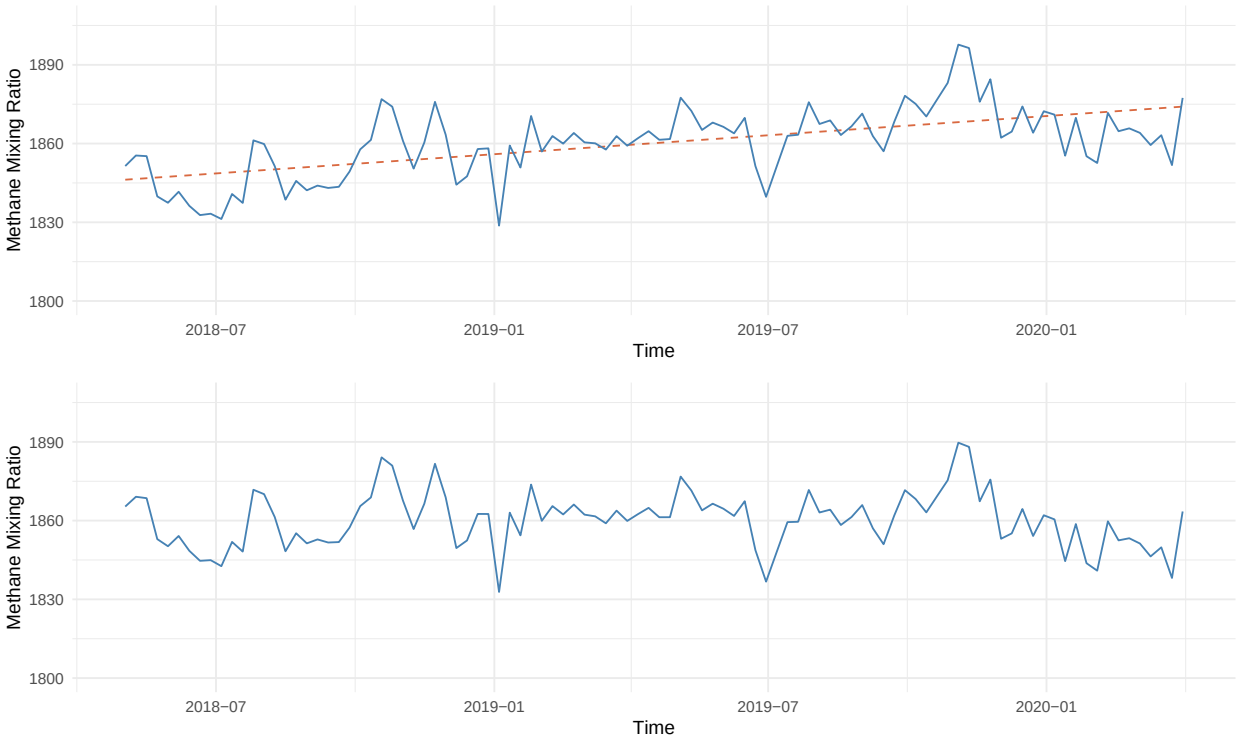


Figure 20: Haynesville: Mean Weekly Methane Mixing Ratio

has the potential to dramatically skew the results. Nonetheless, Figure 21 depicts the frequency spectrum for the Haynesville shale region.

The signals here are much less conclusive than in the spectral analysis of the entire region. No particular signal stands out with any appreciable strength. The two signals marked in red have frequencies of 0.18 and 0.35 months^{-1}, with periods of 5.7 and 2.9 months, respectively. Signals at these frequencies were also singled out in the region at large, but their relative strength here is severely diminished. The presence of many other signals at moderate relative strength is a possible indication of greater variance in methane levels in the Haynesville region over time.

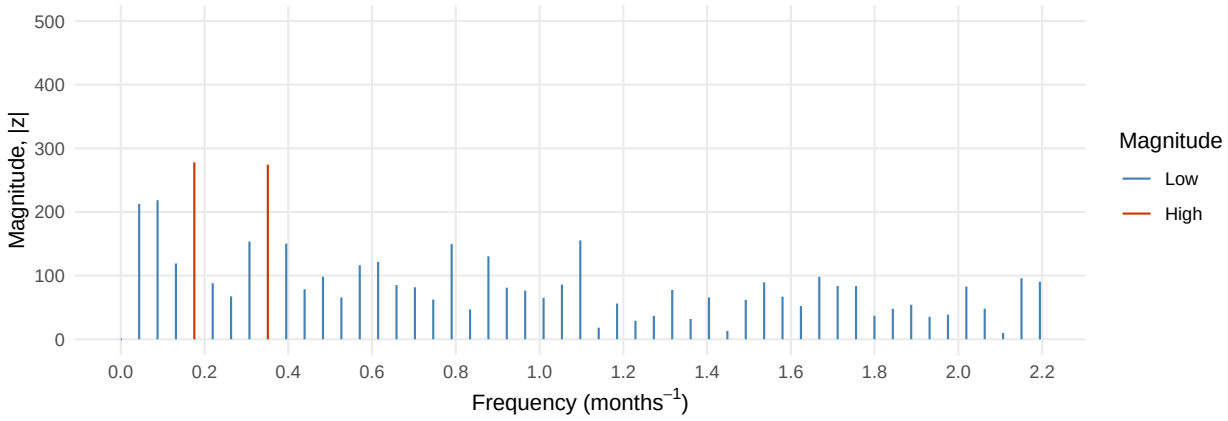


Figure 21: Haynesville: Frequency spectrum for weekly methane mixing ratio

10.2 The Wetlands

Figure 22 contains time-series plots of trended and de-trended mean weekly methane concentrations in the wetlands regions using high-quality data.

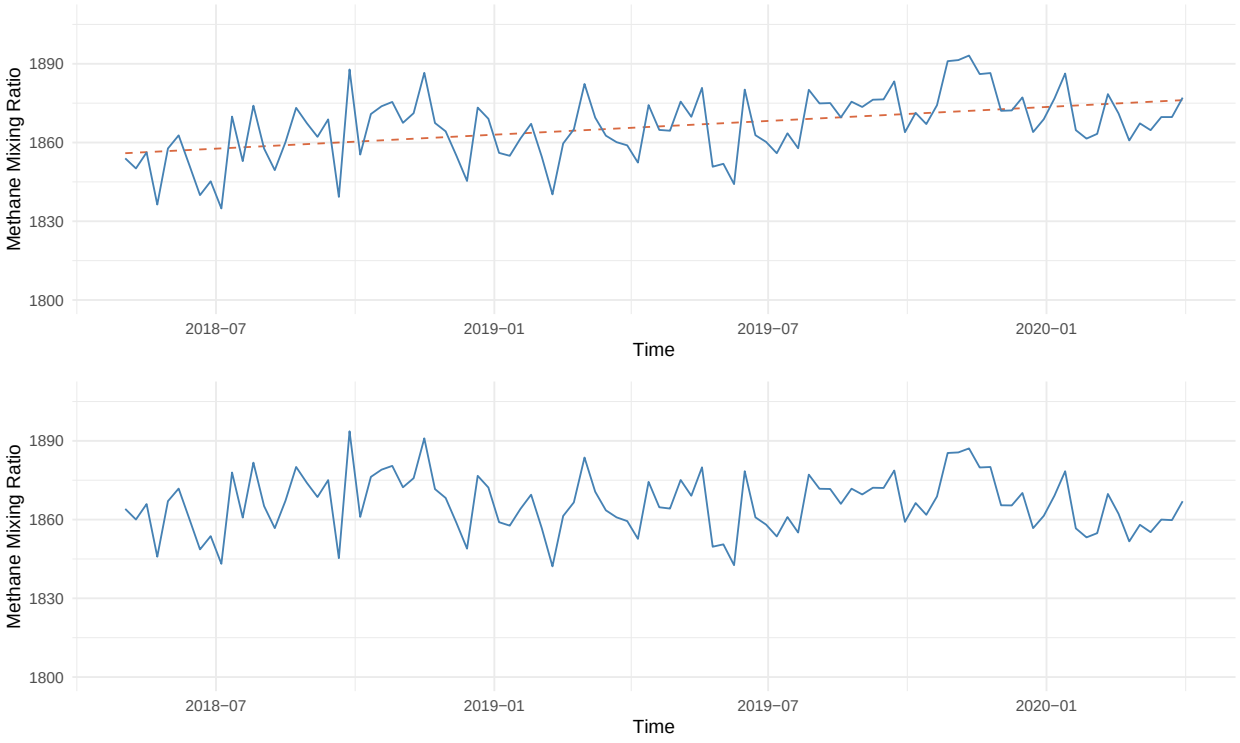


Figure 22: Wetlands: Mean Weekly Methane Mixing Ratio

In this case, 8 of the 100 weeks contained no valid measurements. Figure 23 shows the results of the discrete Fourier transform.

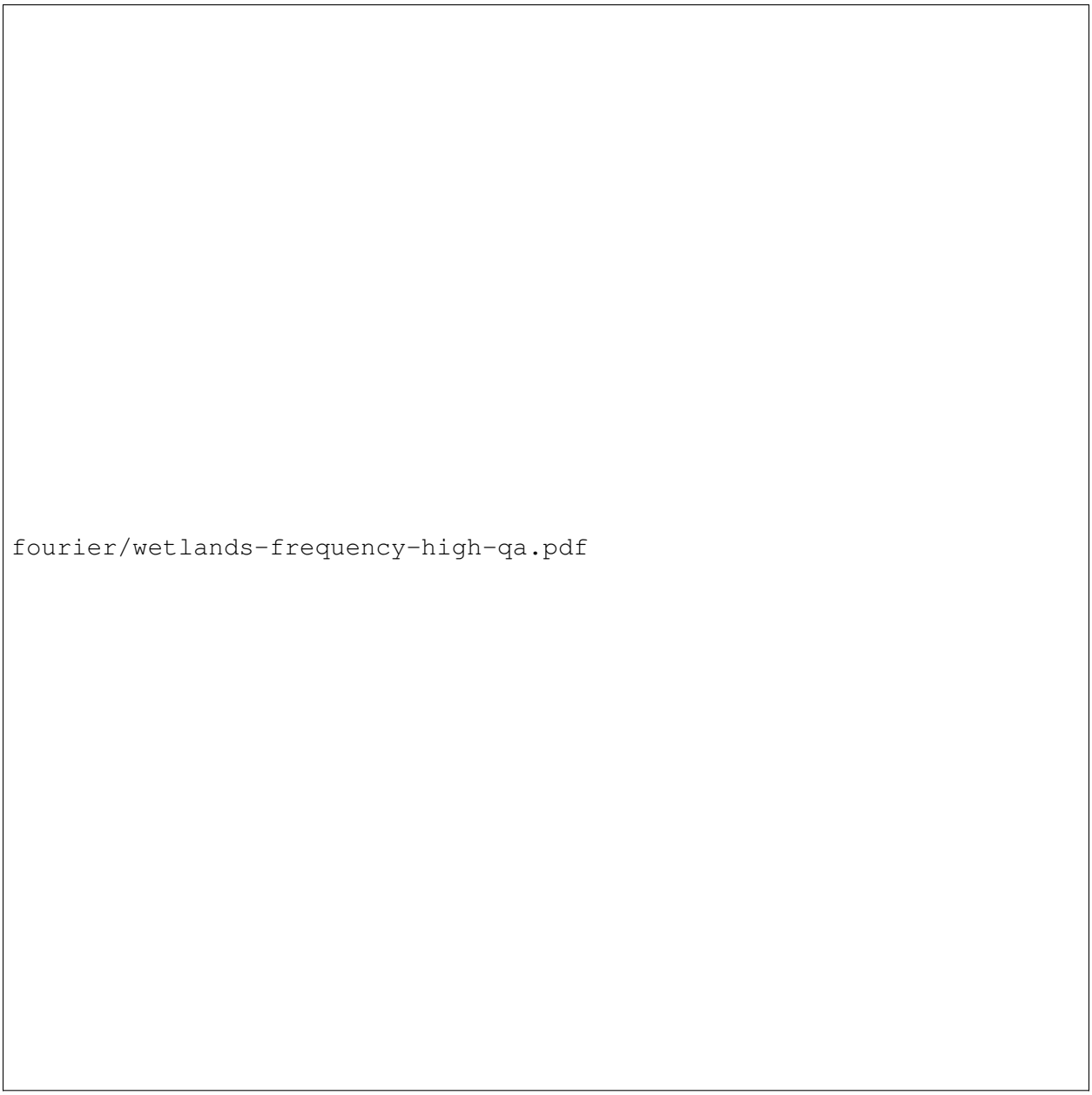


Figure 23: Wetlands: Frequency spectrum for weekly methane mixing ratio

The signals at frequencies of 0.04 and 0.09 months^{-1} stand out here, though, again, there are many other signals with modest strength. These frequencies have associated periods of 23 and 11.5 months, again corresponding to long-term trends in methane concentration. With most of the data in the wetlands regions intact, it is reasonable to conclude that the annual signal exists and is not a result of random noise.

10.3 The Dallas-Fort Worth/Austin Corridor

Figure 24 contains time-series plots of trended and de-trended mean weekly methane concentrations in the Dallas-Fort Worth/Austin region using high-quality data.

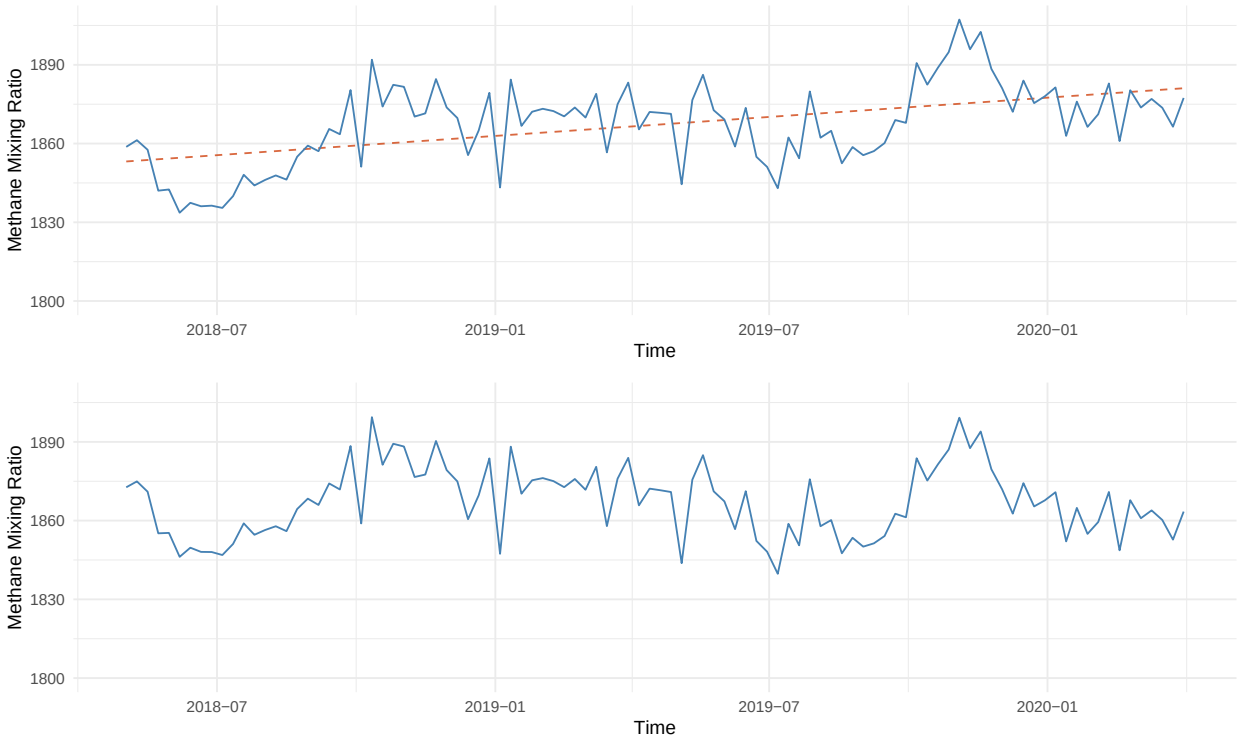


Figure 24: Dallas-Fort Worth: Mean Weekly Methane Mixing Ratio

In this case, 21 of the 100 weeks contained no valid measurements. Figure 25 shows the results of the discrete Fourier transform.

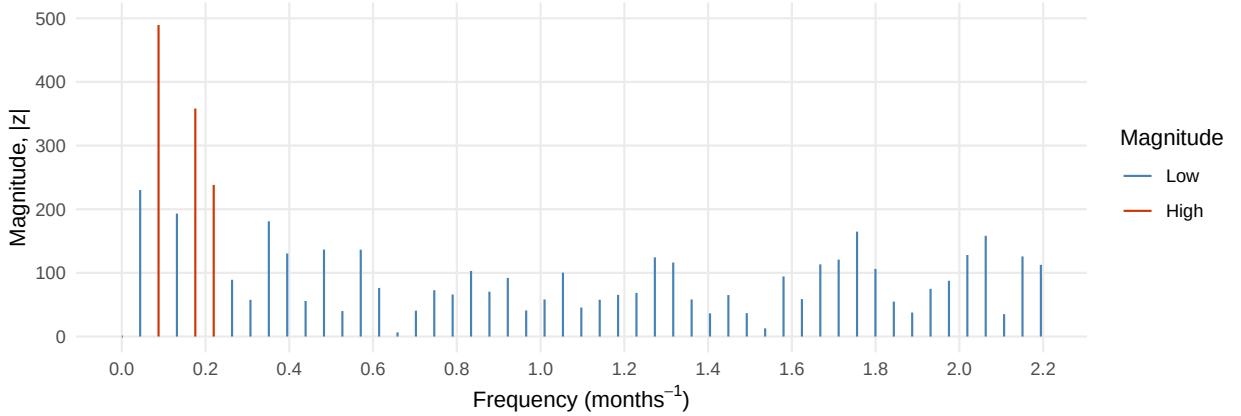


Figure 25: Dallas-Fort Worth: Frequency spectrum for weekly methane mixing ratio

Here, the signals at frequencies of 0.09 and 0.18 months^{-1}, with periods of 11.5 and 5.7 months, are particularly strong, indicating that the previously observed seasonal trend is likely present.

10.4 Analysis

When considering our region at large, there are some very clear and distinct periodic trends. As shown in Figure 19, several long-term trends in methane concentration stand out with notable strength, and it is reasonable to conclude that there exist annual and seasonal fluctuations in methane levels, possibly of natural origin. As to whether these trends exist within our specific regions of interest, it is difficult to say. High-quality data in these regions is sparse, and spectral analysis did not yield any iron-clad results. It is very likely that these trends do exist, but more data and further analysis would be needed to be certain.

The most natural question to ask is: do these trends correspond to natural variation, or does there exist some correlation with artificial sources? Very possibly, the answer is *yes* to both these questions. Confirmation of the latter, however, will require further research.

11 Conclusions

We evaluated the Haynesville and surrounding regions for methane production as methane coming from Haynesville is one of the highest rates in the United States. Instead of evaluating only the Haynesville region, we took a much wider area, encompassing all of Louisiana, East Texas, and Southern Arkansas. This was necessary to develop context about methane production and behavior in surrounding regions, and in doing so, we confirmed previous research that car emissions and microorganisms in wetlands regions contribute to methane.

After breaking the region into many subregions, we discovered which areas contained proportionally high values of methane in Section 3.2. After identifying these regions, we subtracted the base methane levels by identifying wind vectors to isolate methane production in Section 7. We then fit ARIMA models to analyze the behavior of these regions against the Haynesville region in Section 12.2. Finally, after concluding the behavior was similar to that in the other high proportion regions, we fit a linear model in an attempt to correlate natural gas extraction and production to high methane levels through time in Section 9. These results indicated to us, with existing and available data, a correlation cannot conclusively be identified between natural gas extraction and methane levels, a relationship we wanted to establish, as identified in Section 1.2. While we did not conclude this relationship is clearly evident, we believe with more data relating to natural gas extraction, a relationship could plausibly be identified, as a positive, linear relationship appears to exist.

12 Appendix

12.1 Agriculture Findings

We first gathered information on Texas counties [11] and then a census of cows in Texas [3]. From this data, we built a population map from the 2012 data, shown in Figure 26, that should help us get an idea of population densities for Texas cattle.

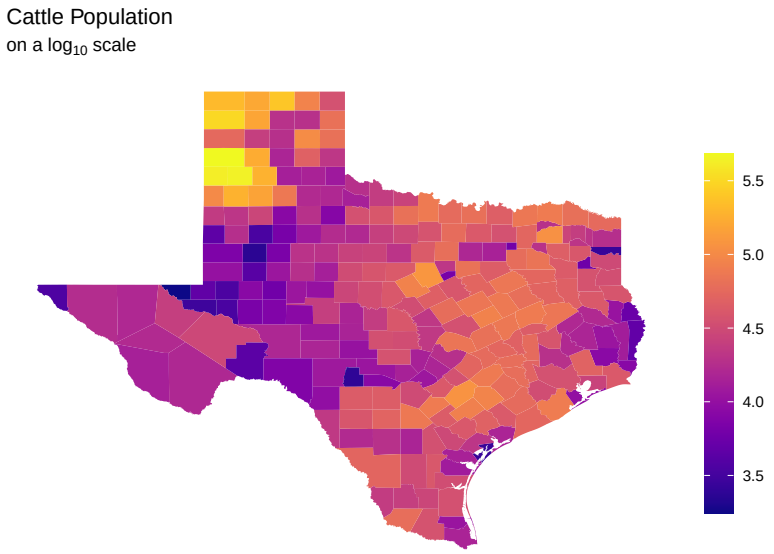


Figure 26: Texas Cattle Log Population Densities

From this map alone, we see no specific area in our region of interest contributing an extremely high relative amount of methane. We conclude, from the visual distribution, agriculture may contribute a significant amount, but nothing anomalous in our region. Furthermore, this data is not recent enough on which to perform statistical inference; therefore, we cease to evaluate agricultural contributions.

12.2 Time Series Analysis

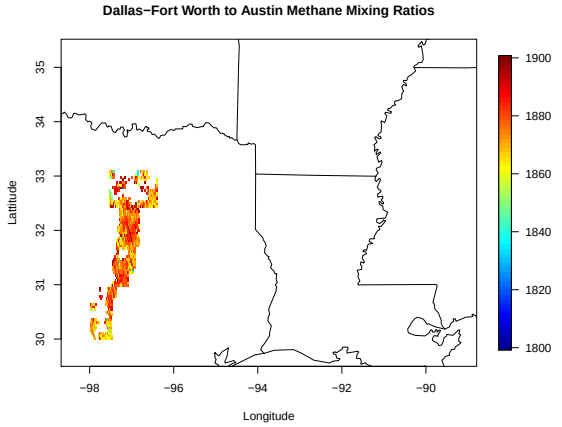
12.2.1 Three Region Analysis

An important feature to look at in these data is the variable time – we can not only see the long-term behavior of multiple regions, but we may also see how these regions behave in relation to each other. It is interesting to know how methane levels have changed through time in our region as well as predict where they might be at some time in the future. The land-water mask ensures we eliminate any bad values that can occur over water and be overlooked by the numerous other data collection precautions in place already. Spatial analysis is a tool that helps identify areas of interest, but as outlined in Section 3.1, spatiotemporal analysis requires knowledge exceeding the scope of this report when many sectors are used; therefore, we need a different tool.

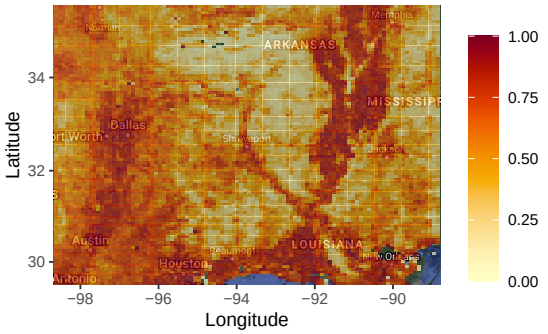
Time series analysis using an Auto-Regressive Integrated Moving Average (ARIMA) model is used to identify and predict trends in time series data inside regions of particular interest. Particular

regions of interest include the stretch along I-35 between Dallas-Fort Worth and Austin, Texas; the Haynesville Shale region (Figure 3); the wetlands area along the Mississippi River, and Red River tributary; and the Mississippi River Delta emptying into the Gulf of Mexico. These regions are well-shown in Figure 10.

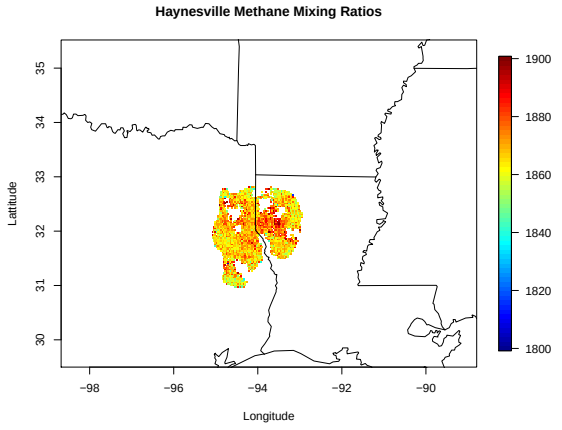
We first start by applying our land-water mask and wind corrections to the data with `qa` values of 1. That data is then separated by regions of interest: DFW-AUS, Haynesville, and Wetlands regions - we do not use data outside these regions. To do this, we used an ad hoc method of box aggregation in lieu of something like Google API shapefiles (except for the Haynesville region, where one was available), brute force, etc. and pieced our data together with small boxes generated with a bounding box tool. The bounding box tool allowed us to trace small, precise rectangles over our region to create a set of geographical boundaries which characterized our regions. [3] The areas generated can be seen in Figure 27 compared to Figure 10:



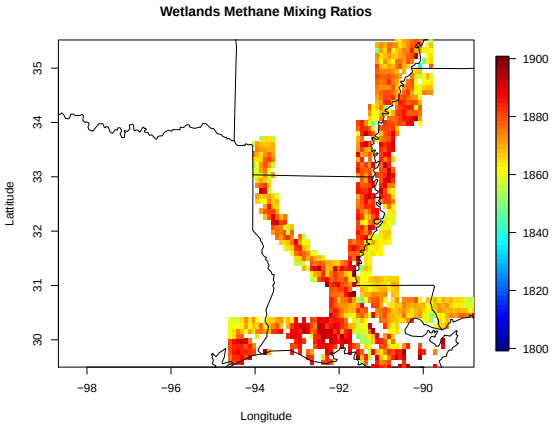
(a) Dallas-Fort Worth to Austin Region



(b) Data with Land-Water Mask



(c) Haynesville Region



(d) Wetlands Region

Figure 27: Three Regions vs. Data with Land-Water Mask

We can see these regions we generated match up very well with areas of high proportions, and thus, interest. We notice the Wetlands regions has the most high-level methane values and the

[3] `https://boundingbox.klokantech.com/`

Haynesville region has the fewest. Some values in these plots are not shown, as they fall outside the plotted range of methane mixing ratios - this is necessary to differentiate the levels in each plot.

To develop the ARIMA model, we take all the high quality data with the mask and wind correction applied, and separate the masked data by region. With the masked data organized by region, we then divide it further into weeks between May 1, 2018 and March 31, 2020. We chose a span of one week because a time period as short as a day does not have enough data frequently and will pick up errors like noise that we want to minimize. Anything longer, and we could miss important trends. In each week in each region, we take the average across the region. If an entire week is missing data, we use the average of the previous week. From this newly organized data, we have a set of data equally spaced by weeks with one entry. We use this to create our ARIMA models, using the `auto.arima()` function which selects hyperparameters on its own, and are shown in Figures 28 and 29:

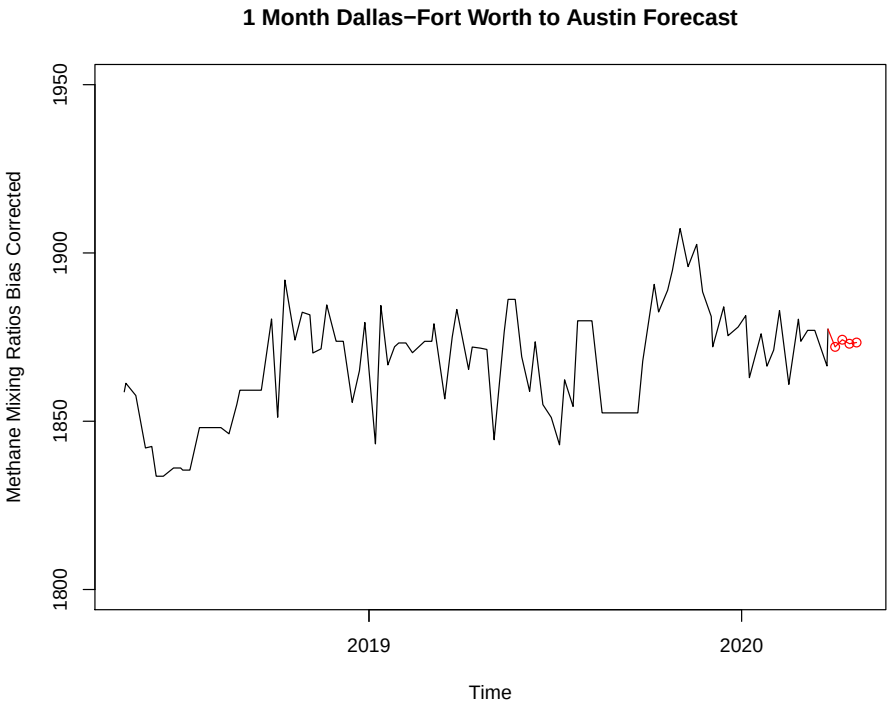


Figure 28: Dallas-Fort Worth to Austin ARIMA(1,1,1) Model with 1 Month Prediction

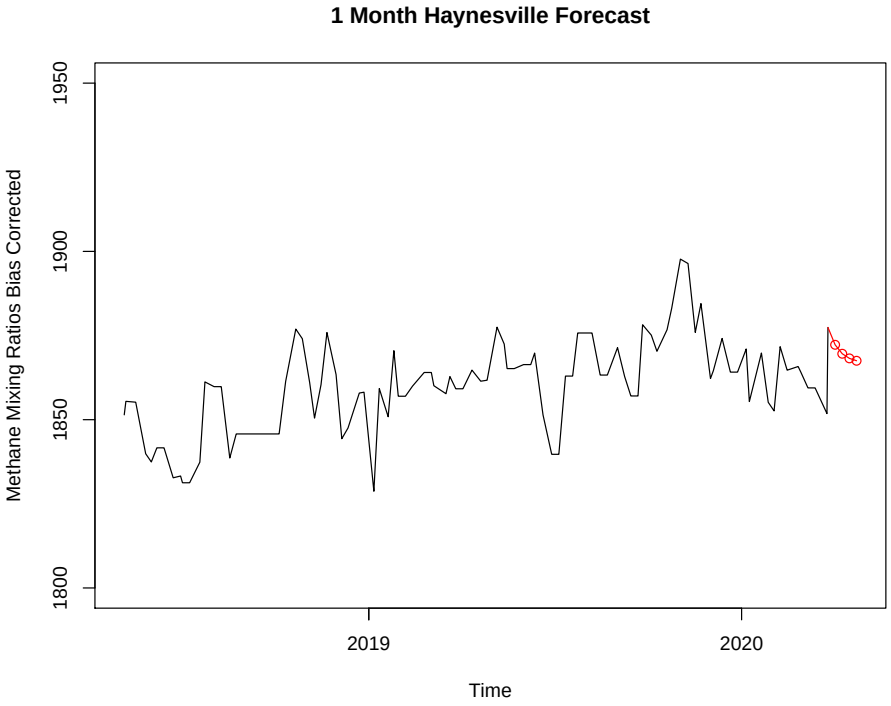


Figure 29: Haynesville ARIMA(2,1,2) Model with 1 Month Prediction

The black lines are the observed values and the red circles are the predicted next four weeks. The Wetlands region generates an initial model recommending the use of 0 previous lags, an indicator that time series analysis is inappropriate here. We notice some fairly obvious things: removing the marginal data left a few sizeable holes which we had to fill in; the next month in each region appears to steadily increase and taper off. While time is likely a useful component, we found little value in the ARIMA models.

13 References

[1] American Society of Agronomy. Unexpected culprit: Wetlands as source of methane. Retrieved Apr. 13, 2020 from https://www.sciencedaily.com/releases/2019/06/190619085703.htm.

[2] Apituley, A., Pedergnana, M., Sneep, M., Veefkind, J. P., Loyola, D., and Hasekamp, O. *Sentinel-5 precursor/TROPOMI Level 2 Product User Manual Methane*, 0.11.6 ed. Royal Netherlands Meteorological Institute, PO Box 201, 3730AE De Bilt, The Netherlands, 6 2017.

[3] Carson, D. Texas counties: Cattle population in 2012. Retrieved Feb. 24, 2020 from http://www.texascounties.net/statistics/cattle2012.htm.

[4] Environmental Defense Fund. Major studies reveal 60% more methane emissions. Retrieved Feb. 13, 2020 from https://www.edf.org/climate/methane-studies.

[5] Environmental Defense Fund. Methane: The other important greenhouse gas. Retrieved Feb. 13, 2020 from https://www.edf.org/climate/methane-other-important-greenhouse-gas.

[6] Environmental Protection Agency. Understanding global warming potentials. *Greenhouse Gas Emissions* (2 2017). Retrieved Feb 3, 2020 from https://www.epa.gov/ghgemissions/understanding-global-warming-potentials.

[7] Green Vehicle Guide. Vehicle emissions. Retrieved Apr. 13, 2020 from https://www.greenvehicleguide.gov.au/pages/Information/VehicleEmissions.

[8] Kahle, D., and Wickham, H. ggmap: Spatial visualization with ggplot2. *The R Journal 5*, 1 (2013), 144–161.

[9] NASA. Modis webpage. Retrieved Mar. 5, 2020 from https://modis.gsfc.nasa.gov/about/.

[10] Office of Fossil Energy. Shale gas. Retrieved Feb. 21, 2020 from https://www.energy.gov/fe/shale-gas-101.

[11] Texas Open Data Portal. Texas counties centroid map. Retrieved Feb. 24, 2020 from https://data.texas.gov/dataset/Texas-Counties-Centroid-Map/ups3-9e8m.

[12] U.S. Energy Information Administration. Haynesville natural gas production is increasing but remains lower than previous peak. Retrieved Apr. 13, 2020 from https://www.eia.gov/todayinenergy/detail.php?id=37033.

[13] U.S. Energy Information Administration. Maps: Oil and gas exploration, resources, and production. *The R Journal* (10 2019).

[14] U.S. Energy Information Administration. Natural gas explained: *Where our natural gas comes from*. Retrieved Feb. 13, 2020 from https://www.eia.gov/energyexplained/natural-gas/where-our-natural-gas-comes-from.php.

[15] Veefkind, J. Tropomi on the esa sentinel-5 precursor: A gmes mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment* (2011), 70–83.