Norwegian School of Economics
Bergen, Spring 2020

Spatial Modelling of Unconventional Wells in the Niobrara Shale Play

A Descriptive, and a Predictive Approach

Vegard Hokstad & Dzenana Tiganj

Supervisor: Jonas Andersson

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

Firstly, we would like to thank our supervisor, Jonas Andersson, for the encouragement and valuable feedback throughout the process of writing this thesis. Furthermore, we would like to thank Artem Abramov at Rystad Energy for providing us with the production- and well-design data, as well as providing us with valuable insight regarding the shale oil industry, and the Niobrara shale play in general. We would like to thank Ketil Hokstad for helping us collect the geological data used for this research. Lastly, we would like to thank Roger Bivand for sharing his expertise regarding spatial data analysis.

Norwegian School of Economics

Bergen, June 2020

Abstract

This research investigates oil production in the modestly studied Niobrara shale play, using data containing information about well-design and production volumes from wells drilled in the period 2011 - 2018. Firstly, machine learning techniques were employed to conduct a descriptive analysis, with the motive of identifying drivers of well-productivity. Models of increasing spatial resolution were applied to isolate the effect of high-grading of geological conditions from well-design choices. The statistical models employed were different Random Forest (RF) configurations, Geographical Random Forest (GRF), and the well-established technique of Regression Kriging (RK). It was found that spatial effects were attributed slightly short of 40% of the relative importance in explaining variations in the first-year production volumes of oil. Further, it was found that models attributed too much importance to well-design variables if spatial effects were not adequately accounted for. It was also found that the more data-driven and less restrictive RF and GRF performed slightly better than the widely recognized RK. Secondly, a predictive analysis was conducted in an attempt at identifying undrilled locations with favorable geology for future drilling. For this part of the research, data containing information about geological conditions were utilized, alongside the production data. It was found that applying RF and GRF yielded quite disappointing results when the task was formulated as a regression problem. However, lowering the precision by rephrasing the task as a classification problem resulted in a RF classifier that outperformed random guessing to some extent. A visual assessment of the classifier's generated heatmap of predictions suggested that the model was highly capable of identifying geological settings associated with the most extreme wells, in terms of productivity.

Keywords – Unconventional Wells, Unconventional Petroleum Resources, Niobrara, Machine Learning, Random Forest, Spatial Data

Contents

List of Figures

List of Tables

1 Introduction

For a long while, it seemed like US oil production had peaked at a level slightly short of 10 million barrels per day in 1970, steadily declining to a level of approximately 5 million barrels per day in 2005 (EIA, 2020a). This development led the US to become highly dependent on imports from countries like Venezuela and Saudi Arabia, which sparked further concerns about oil supplies. It was debated whether a global peak in oil production had been reached - a thesis which gained traction and helped oil prices surge well above $100 per barrel (Rapier, 2017). The situation was quite similar for natural gas production, which also had peaked in the early 1970s, but which had started to recover slowly since the late 1980s (EIA, 2020b). Industry insiders forecasted an energy crisis for the US, but at the same time, several factors helped jumpstart what became to be known as the fracking boom. The fracking boom refers to the vast increase in US oil and gas production, which started with natural gas in 2005, followed by oil in 2008. By 2015, US gas production had risen for 10 years straight, and US oil production had experienced its fastest increase in history. This was facilitated by the combination of hydraulic fracturing and horizontal drilling. Neither were new technologies at the time. The former had been around since the 1940s, while the latter had been around since the 1980s (Rapier, 2017). However, the high oil and gas prices, accompanied by growing demand, helped spark capital investment and further technological innovation. This, along with other factors such as the infrastructure of pipelines, government policy, and land rights ownership, made oil and gas extraction from tight shale rock commercially viable (Manfreda, 2015; Montgomery & O'Sullivan, 2017; Statista, 2020). As a result, the US went from being the world's third largest producer of oil, into becoming the world's leading country in both oil and natural gas production over the past decade (Elliot & Santiago, 2019).

With the last decade being characterized by a sustained fall in oil prices, with further plummeting prices triggered by the Coronavirus outbreak and the concurrent price war between Saudi Arabia and Russia, the shale industry's financial viability is endangered (Markets Insider, 2020; McDonnell, 2020). Thus, smaller margins cause the shale industry's future to be largely dictated by the capacity of operators to improve their productivity (Montgomery & O'Sullivan, 2017). During the early days of the fracking boom, measures towards improving productivity tended to be inconsistent. For instance, Covert (2015)

conducted an empirical analysis of productivity gains of wells between 2005 and 2011. He found that companies tended to learn slowly from their operations and incrementally improved technology. Interestingly, he also found that they often failed to learn from competitors and that a focus on reducing costs led to a suboptimal lack of experimentation. In 2012, the growth in new oil and gas rigs started slowing off. However, production kept on increasing despite this, indicating that oil and gas producers began getting more out of less (EIA, 2020a; Trading Economics, 2020). Much of this increase in productivity was attributed to new trends in well-design such as longer lateral lengths, and increased volumes of fracking fluid and proppant mass (Gold, 2014). Given the vast interplay of factors that influence the productivity of operators, data analytics, and simulation of variations in well-productivity based on drilling decisions have become of interest for researchers. Statistical analysis and data-driven modeling are considered to potentially be driving future decision-making processes in the industry. With increasing data availability, researchers aspire to support decision-makers with drilling decisions and processes (Fu, 2019). This research will focus on the following two objectives: Firstly, a selection of machine learning methods will be used to establish an understanding of which factors have been driving the increase in well-productivity through the last decade. Secondly, machine learning methods will be employed with the aim of predicting oil production volumes at undrilled locations, based on several geological variables. The analyses for both parts are conducted with the programming language R (R Core Team, 2020). The next subsection will present further details of the study scope and aim of this research.

1.1 Study Scope and Aim

The Permian (Wolfcamp and Bone Spring) and Williston (Bakken and Three Forks) basins are the central areas of unconventional oil production in the United States and have to a large extent put the country in a leading position within the global oil industry. The basins' associated shale plays are in parenthesis (EIA, 2019; Reuters, 2019). These fields, among others, have been widely studied and explored compared to plays like the Niobrara of the Denver-Julesburg Basin, forming a research gap. To exemplify, Google Scholar generates 14,100 oil and gas-related article hits for "Niobrara" compared to, for instance, 78,400 and 174,000 for "Bakken" and "Permian", respectively. The search terms were chosen based on whether the area is usually referred to with the name of the basin or the shale play itself.

Nevertheless, analysts suggest that there is a lot of untapped potential in the Niobrara play, with estimates of oil richness being higher than in the Bakken play (DiLallo, 2018; Hart Energy, 2011). The realization of these estimates is, however dependent on drilling decisions made by the operators. The abovementioned characteristics make the Niobrara play an interesting study area, and was therefore chosen as the area of interest for this research. It is located in parts of Colorado, Kansas, Nebraska, and Wyoming (Speight, 2017), and has been in production since 2006 (Ahmed & Meehan, 2016).

This research aimed to investigate the modestly studied Niobrara shale play and to provide potential decision-makers with relevant insight that may support future drilling decisions. This was conducted through analyzing the data at hand from two perspectives:

1. Descriptive perspective:

   For the first part, statistical models were applied to historical well-data, in order to investigate the relative impact of well-design and reservoir quality, concerning the first-year oil production of unconventional wells. For this part, well-location was used as a proxy for reservoir quality. The effect of controlling for location was assessed by constructing models of increasing spatial resolution. Random Forest (RF) was chosen as the basis for modeling, and was adjusted to more sophisticatedly account for spatial effects in a stepwise manner. This is described in more detail in Chapter 5.3.1. Additionally, a variant of RF specifically designed for spatial data, the Geographic Random Forest (GRF), was applied. This extension in terms of spatial resolution was considered most sensible with a regular RF as the baseline. Moreover, RF was chosen since it makes few assumptions about the input variables (Lovelace, Nowosad, & Muenchow, 2019), and is capable of accounting for nonlinearities in the data (Hengl, Nussbaum, Wright, Heuvelink, & Gräler, 2018). This makes it less restrictive than, for instance, regular linear regression models (Molnar, 2019). Additionally, Regression Kriging (RK) was also applied in order to serve as a benchmark for comparison of this study's data-driven approach to more widely recognized, but more manual and less comprehensive approaches (Hengl et al., 2018). Properly assessing the impact of spatial effects is essential in order to make reliable forecasts about future production levels, as well as obtaining an understanding of the resource economics at play.

2. Predictive perspective:

   Subsequent to investigating drivers of well-productivity, this research aimed to apply machine learning techniques to identify favorable locations for future drilling. For this, a RF and GRF were trained to predict first-year production volumes of oil under varying geological conditions. Hence, the models were applied to data containing information about historical production volumes of wells and associated geological conditions at their location. Given the nature of this task, a framework dedicated to the development of spatial prediction models was employed. As will be presented in Chapter 2, previous studies have shown that modeling production volumes based on geological data has proven to be difficult. Thus, the second part applied a more organic and experimental approach to see if historical production data along geological variables could be used to identify favorable geological settings. If successful, such models may be used for initial exploration purposes, to decide whether or not to allocate resources for further investigation of an area, and eventually acreage acquisition.

The next subsection will introduce the concept of hydraulic fracturing to deepen the understanding of the studied industry and previous research. The remainder of the thesis is structured in the following way: Chapter 3 briefly describes the different datasets used for the analyses. Chapter 4 presents the applied methodology, describing special properties of spatial data, the different statistical models employed, metrics for evaluating model performance, and the model development framework. Chapter 5 will elaborate on pre-processing steps conducted prior to the analysis, considerations regarding data partitioning, the different model configurations, as well as the process of tuning the models. In Chapter 6, the results of the two parts will be presented, before being more thoroughly discussed in Chapter 7. Chapter 8 will conclude the research and summarize its contribution to existing literature.

1.2 Introduction to Hydraulic Fracturing

The following section will provide an introduction to the process of hydraulic fracturing, hereby referred to as fracking. The principle of fracking involves cracking tight rock formations open and propping the fissures, in order to allow oil and gas to flow through

the borehole (PSAC, n.d.). This technique is used for extraction of oil and natural gas from so-called unconventional reservoirs. Simply put, conventional reservoirs are focused sub-surface accumulations of hydrocarbons in sandstones or carbonates, sourced from organic-rich shale. In unconventional reservoirs, the hydrocarbons are produced from the source rock itself (Government of British Columbia, n.d.). Thus, a major difference is the permeability, which is a substance's ability to transmit fluids (Schlumberger, n.d.). Due to the low permeability, hydrocarbons cannot be extracted in an economically viable way through traditional vertical wellbores. Instead, unconventional reservoirs require horizontally drilled wells followed by fracking, in order to achieve viable production (Government of British Columbia, n.d.). From the surface, a well is drilled vertically until right before the depth of the selected shale formation is reached. At this point, it curves in order to be drilled horizontally through the shale formation. The length of this horizontal section is called the lateral length, and the longer this is, the greater the contact area with the formation (Statoil, 2013). After being drilled, the well must undergo completion before it may be fracked. This involves running steel pipes, so-called casing, down the inside of the drilled well, and permanently setting it in place by filling the gap between the casing and the rock with cement (Rigzone, n.d.). Fracking commences at the farthest end of the well and continues stagewise towards the vertical section. The number of fracking stages required depends on the lateral length. Most often, a so-called perforating gun is brought down into the well and used to fire small explosive charges in order to open the wellbore to the shale formation. After the wellbore is perforated, a mixture commonly referred to as fracking fluid is pumped into the well at a pressure high enough to create tiny cracks in the rock formation. This mixture consists of water, proppants, and chemicals. Once a stage is adequately fracked, the pressure is relieved, and the cracks begin to close. The proppant, which usually is sand or ceramic pellets, props open the cracks, which provide the hydrocarbons with a flow path into the wellbore. Once the fluid is pumped away, the stage is plugged, and the fracking process is repeated for the remaining stages that constitute the horizontal section of the well. At last, the plugs are drilled out, allowing the oil or natural gas to flow up the wellbore (PSAC, n.d.; SM Energy Company, 2015)

2 Literature Review

This part of the thesis will present findings from previous research that have been central to the various choices made throughout this research. This involves the choice of well-design variables, the choice of statistical models, and choices related to modeling of spatial effects. At last, findings of relevance for decisions regarding the selection of geological variables, and performance assessment of spatial prediction models are presented.

Previous research modelled the potential behavior of shale reservoirs based on the premise that the formation properties were geologically homogeneous, and focused on technological features (Lolon et al., 2016; Zhou, Kleit, Wang, & Dilmore, 2014). Accordingly, a broad consensus has been established that well-design variables are driving forces of production volumes. In particular, an increase in proppant mass and fracking fluid was notably followed by an increase in the production levels of unconventional wells (Curtis & Montalbano, 2017). Along with other important factors, such as the lateral length (Esmaili & Mohaghegh, 2016; Mohaghegh, Gaskari, & Maysami, 2017; Montgomery & O'Sullivan, 2017), these variables have shown to capture variability in shale oil production volumes. A study by Lolon et al. (2016), applied machine learning methods, namely Random Forest (RF) and Gradient Boosting Machine (GBM), to model well-productivity based on well-design variables. Their best result was provided by RF and established further consensus that fracking fluid and proppant mass are strong predictors of well-productivity. Wang and Chen (2019), provide further support for the power of RF in modeling well-productivity and identified influential design variables to be proppant, lateral length and fracking fluid. Notably, these two studies do not only establish consensus for the predictive power of well-design variables, but also imply the strength of RF as a method to model productivity of unconventional wells. RF has been widely used for various applications within the petroleum industry (Attanasi, Freeman, & Coburn, 2020; Bhattacharya & Mishra, 2018; Lolon et al., 2016; Luo, Tian, Bychina, & Ehlig-Economides, 2018; Wang & Chen, 2019). The findings from the abovementioned research motivated the choice of well-design variables and employment of the RF algorithm, for this research.

While earlier studies assumed geological homogeneity, the shale formation is known to exhibit heterogeneity in its properties (Dong, Holditch, & McVay, 2013; Satter & Iqbal,

2015). The natural presence of fractures and other formation properties may be estimated, but are never certain. Moreover, rock properties change horizontally and vertically, which advocates for the uncertainty associated with predicting the performance of unconventional wells. Along with the extremely low permeability and the characteristic heterogeneity of shale, it is argued that it is difficult to estimate reliable production volumes based on geological conditions (Satter & Iqbal, 2015; Xie, Lee, Wen, & Wang, 2013). Despite this uncertainty, a convergence of drilling efforts to areas with favorable geological conditions has been observed (Rystad Energy, 2017a, 2017c). The reservoir capacity is considered to be spatially dependent and the rock properties are highly variable at small spatial intervals (Clarkson, Jensen, & Chipperfield, 2012), indicating that spatial variation may be captured by using geolocation variables as a proxy for geological properties. A study conducted by Montgomery and O'Sullivan (2017), with wells drilled between 2012 and 2015 in the Bakken play, found that spatial effects drove production levels to nearly the same extent as well-design choices. They accounted for spatial dependency of geological properties by using location as a proxy for reservoir quality. Five regression models of increasing spatial resolution were applied. They found that failing to incorporate spatial dependencies can introduce a substantial bias, which may lead to overestimating the impact of technology on well-productivity. Their results from Regression Kriging and Spatial Error Modelling concluded that *high-grading* had an equally strong effect on productivity as well-design variables such as lateral length, fracking fluid and proppant mass. The term "high-grading" refers to the tendency of focusing drilling efforts to locations with the most favorable geology. A study conducted by Rystad Energy (2019) found that spatial effects were of significance for explaining well-productivity, and more in the long run than the short run. They used what Bivand, Pebesma, and Gomez-Rubio (2008) refers to as a trend surface analysis, which is a linear regression where polynomials of geographical coordinates are included as predictors. It was found that under an aggregation period of 3 months, spatial effects explained 22 percent of variability in production volumes. For aggregation periods of 6, 12 and 24 months, spatial effects explained 26, 28 and 29 percent of the variability. Similar studies have been conducted by Zhong, Schuetter, Mishra, Lafollette, et al. (2015) and Izadi, Zhong, LaFollette, et al. (2013) who have, respectively, modelled well-productivity in the Wolfcamp and Bakken shale plays, with help of well-design and geolocation variables. They both found location, thus geological conditions, to be an

important driver of production output. Additionally, Zhong et al. (2015) found that RF performed best of all the nonlinear- (Support Vector Machine (SVM), RF, GBM) and linear (Ordinary Least Squares (OLS)) methods employed. The findings from these studies motivated the choice of employing geolocation variables as a proxy for geological conditions, for the first part of this research. Further, it inspired the choice of model configurations employed in part 1, as well as employing RK as a benchmark model.

The abovementioned studies have proven that geolocation variables are well suited for identifying the effects of high-grading in historical well-data. However, Meyer, Reudenbach, Wöllauer, and Nauss (2019) present that using such predictors can lead to considerable overfitting when the motive is to generate predictions beyond locations used for training the models. They state that such predictors, that are highly spatially autocorrelated, tend to be misinterpreted by algorithms in a way which makes them great at reproducing the training data, but bad at generating predictions beyond it. This suggested that the attempt at predicting into undrilled acreage required using predictors that are more directly determinant of the oil production. Since the aforementioned well-design variables represent information that is not available until a well has been drilled, this task required using geological variables that influence oil generation. It is well-established that oil generation happens in the temperature window of roughly 60 to 130°C, with peak generation at approximately 90°C (Allen & Allen, 2013). Some studies have incorporated this thermal maturity by using the depth of the well's vertical section as a proxy for temperature, and have found it to be among the most important variables for predicting production volumes (Luo et al., 2018; Schuetter, Mishra, Zhong, & LaFollette, 2015). Luo et al. (2018) and Amaechi, Ikpeka, Xianlin, and Ugwu (2019) studied the interplay of geological properties and technological factors and found reservoir thickness (also referred to as isopach) to be one of the key settings of reservoir quality and indicators of sweet spots. It should be mentioned that the latter study focused on gas production and not oil. Other geological variables that have been identified as important predictors are permeability and porosity (Li & Han, 2017; Luo et al., 2018). At last, Meyer et al. (2019) and Lovelace et al. (2019) present an important finding for analyses aiming at generating spatial predictions beyond observed data. They highlight the importance of accounting for spatial proximity of observations when partitioning data for assessing model performance through cross-validation. This is in order to avoid overoptimistic assessments of model performance

caused by *spatial autocorrelation*, a concept that will be explained in Chapter 4.1. The abovementioned studies' findings motivated the exclusion of geolocation variables and inclusion of certain geological variables when the motive was predicting into undrilled acreage in part 2. It also motivated the employment of proper techniques for performance assessment of spatial prediction models.

3 Data

Throughout this research, three different data sets were used for analyzing oil production in the Niobrara shale play. The first dataset, hereafter referred to as the production data, contained monthly time series data regarding different well productivity metrics, such as the oil production in number of barrels (bbl) and gas production in thousand cubic feets (MCf). All the observations were labelled with an identification code for each well. A second dataset, the well-design data, contained data regarding design properties of the wells in the area. Examples included the amount of proppants and fracking fluid used, lateral lengths, and the name of each well's operator. The dataset also contained columns for each well's identification number, as well as different geolocation variables. All this data were from wells drilled in the period 2011 through August 2019. The third dataset, hereafter referred to as the geological data, contained data regarding geological properties, such as reservoir thickness and -temperature, for the area of interest. Each of these observations were tied to their own unique location, through pairs of longitude and latitude points.

3.1 Production Data and Choice of Response Variable

An important first step to consider is to choose a convenient metric of well-productivity. The production data were collected from Rystad Energy's ShaleWellCube, and contained monthly production volumes for each of the wells, labeled by their identification number. Additionally, it contained a variable representing the well's number of active days for each month. The available variables for well-productivity are summarized in Table 3.1.

Table 3.1

Well-productivity variables.

Variable name	Description
OilProduction_Bbl	Barrels of oil produced
GrossGasProduction_MCf	Thousand cubic feets of natural gas liquids (NGL) produced
TotHydrocarbonProd_6_to_1_Boe	Two-stream production (oil and NGL), where gas volumes are converted to Boe with a factor of 6 MCf to 1 Bbl
TotHydrocarbonProd_20_to_1_Boe	Two-stream production (oil and NGL), where gas volumes are converted to the economic value of 1 Boe with a factor of 20 MCf to 1 Bbl

Source:*Rystad Energy ShaleWellCube*

For this study, *OilProduction_Bbl* was chosen for constructing the response variable. The reason for this is that natural gas may be perceived as just a less valuable byproduct of crude oil production (DiSavino, 2019). The gas may be captured and brought to the market through pipelines, but when they are not available, operators often end up getting rid of the gas through *flaring* (igniting the gas) or *venting* (releasing it directly into the air) in order to continue the oil production. Even though it has not happened with natural gas from Niobrara, local natural gas prices in the Permian Basin in Texas turned negative multiple times during 2019 (Collins & Adams-Heard, 2019). With this in mind, oil production volumes were considered a more interesting response variable than the total hydrocarbon production.

Further, a convenient aggregation period for the response variable had to be decided upon. The well-productivity typically tends to peak sometime during the first couple of months, before starting to decrease steadily and converge (Montgomery & O'Sullivan, 2017). Due to this, a longer time frame is usually believed to be a more representative measure of productivity, as it is to a greater extent influenced by a combination of well-design, rock quality, and completion intensity, rather than the individual operator's completion technique (Rystad Energy, 2017a). Previous research has found the cumulative first-year

production to be indicative of the productivity peak and how rapidly it declines, causing it to serve as a good indicator for the *estimated ultimate recovery* (EUR) (Montgomery & O'Sullivan, 2017). Further, studies have found correlations in the range of 0.70-0.85 between first-year production and ultimate recovery (Rystad Energy, 2017a). Based on this, the 12-month cumulative oil production was chosen as a metric for the response variable. By choosing this aggregation period, 866 wells had to be omitted from the dataset, since they had not yet been active for 12 months at the time the data were gathered. A drawback related to this is that the most recent data could not be utilized. Thus, the most recent observations used for the analysis were from November 2018. However, since previous research suggests that first-year production is more representative of the EUR than shorter aggregation periods, it was considered a more interesting metric for the response variable. It was contemplated that this justified the cost of having to remove more observations.

3.2 Well-design Data

The well-design data were also collected from Rystad Energy's ShaleWellCube. This dataset consisted of 7439 observations and 27 variables in its original form, but only a handful of these variables were utilized for constructing the models used for the analysis. These variables are presented in Table 3.2 and were considered since they are under the control of the operator, or simply longitude and latitude points serving as a reference to the geographic location. Further, as presented in Chapter 2, their relevance has been established through previous research. Limiting the scope to only these features also allowed for a better comparison of the results of this research's first part to the findings from Montgomery and O'Sullivan (2017), especially. Examples of variables that were excluded from the analysis are metadata like the well's name and its identification code. Further, there was a total of seven different variables related to geolocation, which naturally was almost perfectly correlated. However, some of these variables were kept in the dataset for pre-processing purposes, as will be described in Chapter 5.1.3. Concerning the geolocation variables that were utilized from this dataset, the longitude and latitude points had a precision of 4 and 5 decimals, respectively. This implies a resolution of approximately 8.5 meters in terms of longitude and 1.1 meters in terms of latitude (Veness, n.d.). In fact, the dataset contained longitude and latitude points for both the wellhead

and the bottomhole. The wellhead is where the well penetrates the surface, while the bottomhole is the end of the wellbore. It was chosen to use the bottomhole coordinates as location reference since they were considered more representative of the target formation's location. In the following, these will just be referred to as longitude and latitude.

Table 3.2

Well-design variables.

Variable name	Description
lat_length	Length (feet) of the horizontal section of the well
proppant	Amount of proppant mass (pounds) per foot of lateral length
frac_fluid	Amount of fracking fluid (barrels) per foot of lateral length
longitude	Decimal degrees with respect to the Prime Meridian, 4 decimals.
latitude	Decimal degrees with respect to Equator, 5 decimals.

Source:*Rystad Energy ShaleWellCube*

3.3 Geological Data

The geological data consisted of 31,215 observations and 9 variables, and was collected by digitizing maps from various published studies. This is outlined in the end of this section. Geographic coordinates, longitude and latitude, tied each of the observations with the remaining seven variables to a unique location. In other words, each predictor represented a spatial property at a specific location, given by the pair of longitude and latitude points. In this dataset, the longitude and latitude points had a precision of two decimal points. This means that each location was approximately 0.85 kilometers apart in terms of longitude, and 1.1 kilometers apart in terms of latitude (Veness, n.d.). Table 3.3 provides a brief explanation of the different predictors in the geological data.

Table 3.3

Geological variables.

Variable name	Description
isopach	The thickness of the formation (meters)
topNio_msl	Elevation at the top of the formation, relative to sea level (meters)
topNio_surf	Depth from the surface to the top of the formation (meters)
surf_elev	Elevation at the surface, relative to sea level (meters)
slope_angle	The slope at the top of the formation, measured in degrees
temp_gradient	The increase in temperature per kilometer of increased depth
topNio_temp	The temperature at the top of the formation
longitude	Decimal degrees with respect to the Prime Meridian. Precision of 2 decimals.
latitude	Decimal degrees with respect to Equator. Precision of 2 decimals.

The *isopach* data were collected from Longman, Luneau, and Landon (1998), while the *topNio_msl* data were collected from Han et al. (2019). The surface topography, *surf_elev*, was gathered from the GEBCO 2019 public database (GEBCO, 2019). *topNio_surf* was computed as the difference between *surf_elev* and *topNio_msl*, and the *slope_angle* was computed by differentiation of the *topNio_msl* variable. The *temp_gradient* was collected from Thul (2012), and *topNio_temp* was computed as:

$$T(z) = T_0 + g(z_0 - z) \tag{3.1}$$

Here, T_0 represents the mean annual surface temperature of 12 °C in the Denver, Colorado area. g represents the *temp_gradient*, and $z_0 - z$ represents *topNio_surf*.

3.4 Exploratory Analysis

As mentioned earlier, vulnerable profit margins put pressure on operators to improve the economics of extraction by increasing productivity. Figure 3.1 visualizes the mean first-year production volume of oil over time, for the wells used in this study. The production volumes are the volumes produced during the first year of activity, for wells initiated in the quarter given by the x-axis. The figure shows a clear upward trend indicating productivity improvements over time. The mean first-year production was approximately 37,000 bbl for wells initiated in Q1-2011, and approximately 112,000 bbl for wells initiated in the peak quarter, Q4-2016.

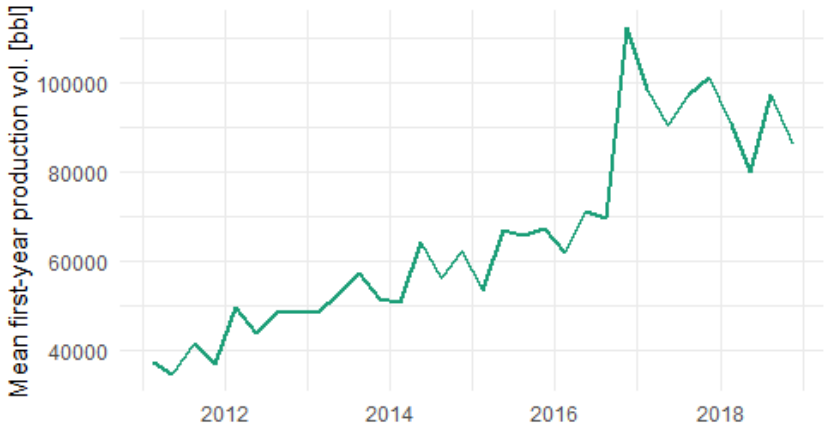

Figure 3.1. Mean first-year production volume over time.

As outlined in Chapter 2, there is a broad consensus that one of the main drivers of this improvement in well-productivity is the upscaling of well-design parameters such as fracking fluids, proppant mass and increased lateral lengths. Figure 3.2 visualizes the quarterly development of production volumes alongside different well-design variables. All variables are scaled and centered quarterly figures. It can be seen that the different curves tend to follow a somewhat similar pattern. For instance, the distinct peak in first-year

production volumes of Q4-2016 was also the peak month for both proppant mass and fracking fluid. The observed pattern would suggest that increases in productivity could be achieved by maintaining or increasing the magnitude of these well-design variables.

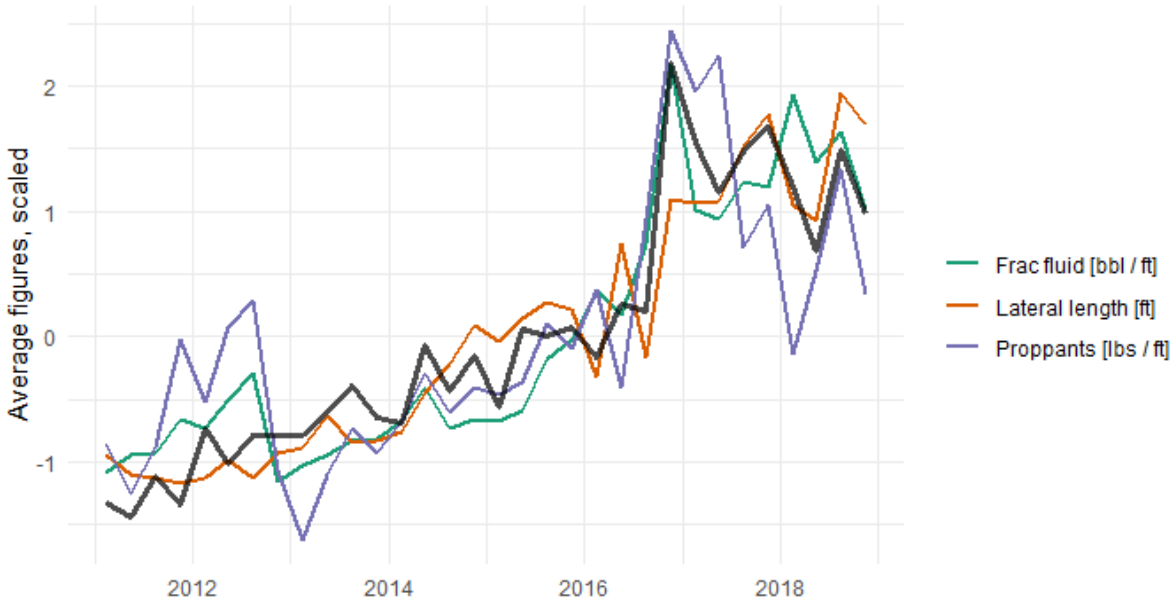

Figure 3.2. Mean levels of well-design along mean first-year production volume, over time. All figures are scaled and centered.

However, another important aspect is revealed by plotting the number of new wells initialized per quarter, and their associated mean first-year production volumes (Figure 3.3). The peak quarter in terms of productivity, is also the quarter where a substantial dip in the initialization of new wells occurs. Thus, there was relatively few wells behind this distinct spike in productivity. A reasonable cause might be that during a quarter where investment conditions are non-lucrative, for instance characterized by high interest rates or a bad oil price outlook, wells are picked and initiated more selectively than in quarters where these conditions are more advantageous. This might cause operators to only initialize new wells at locations where the geological conditions are most favorable, in order to ensure economically sustainable operations during times when margins are under pressure. This is the so-called high-grading practice that was introduced as a driver of well-productivity in Chapter 2.

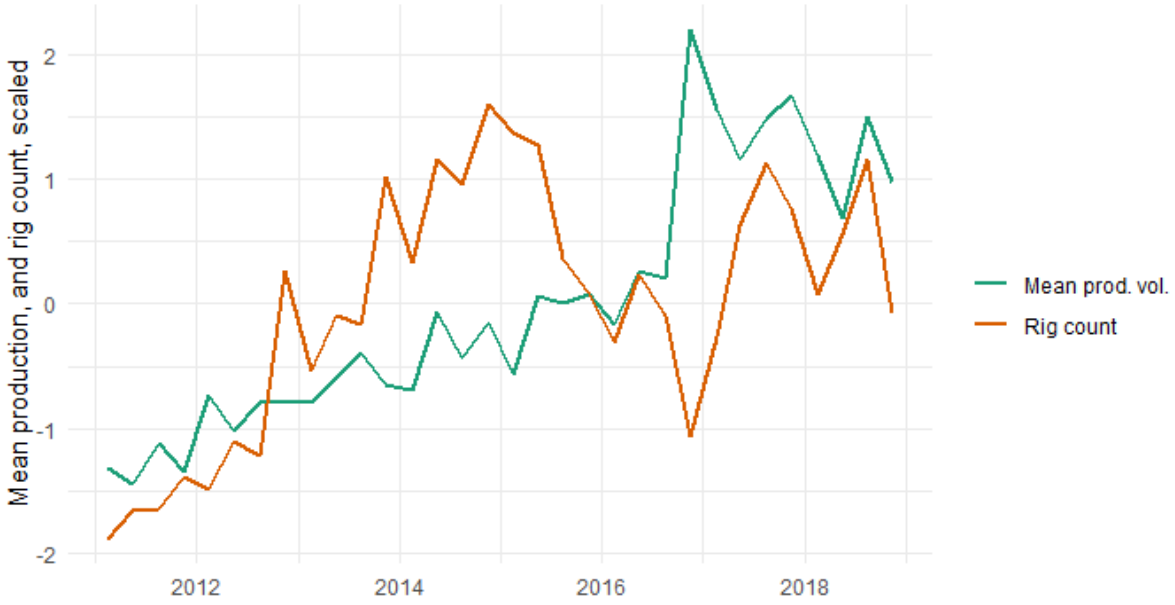

Figure 3.3. Mean first-year production volume along number of new wells initiated, per quarter.

Based on the discussion above, the underlying data used for this research appeared to show similar characteristics to data from other shale plays, used in previous research.

4 Methodology

This part of the thesis will begin by presenting some of the special properties of spatial data, before explaining the intuition behind the statistical models employed for the analyses. Further, the different performance metrics used for assessing model performance are presented, as well as the model development framework.

4.1 Spatial Data

Spatial data refers to data with an associated spatial reference, thus containing information about a specific location on the surface of Earth. This spatial reference may, for instance, be pairs of longitude and latitude decimal degree values with respect to the Prime Meridian, the 0° of longitude at Greenwich in the UK, and Equator, the 0° of latitude (Bivand, Pebesma, & Gomez-Rubio, 2013). This is how the data used for this study was

spatially referenced. Waldo Tobler formulated the first law of geography as "everything is related to everything else, but near things are more related than distant things" (Miller, 2004). This introduces a crucial property of spatial data, in which observations that are spatially near each other tend to be more related than distant observations. This spatial pattern, which is known as spatial autocorrelation, may contain useful information about unobserved factors that influence the variable of interest. If a model is misspecified and not capable of taking this spatially patterned information into account, it may result in spatially autocorrelated residuals, indicating biased models (Bivand et al., 2013). Spatial dependencies may also complicate the use of machine learning techniques because spatial autocorrelation can lead to information leakage between training and test sets. This may further yield biased performance estimates of spatial prediction models, if designated techniques for spatial cross-validation are not applied (Lovelace et al., 2019; Meyer et al., 2019). Since well-productivity is influenced by the geological properties at its location, the data used for this research was spatially dependent. This raised the need for covering the topic of measuring spatial autocorrelation in model residuals (Chapter 4.3.2) and data partitioning with spatial data (Chapter 5.2)

4.2 Statistical Models

4.2.1 Kriging

Kriging is a geostatistical technique that allows accounting for spatial autocorrelation when predicting a response variable. If variables are spatially autocorrelated, then the location at which the observations were measured may explain some of the variability (Goovaerts, 2006). As such, the residual error ϵ in equation 4.1, that exhibits spatial autocorrelation, can explain some variance in the spatially dependent data such that:

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0,\ \Omega) \tag{4.1}$$

This spatial autocorrelation, as represented by Ω in the error term, can be modelled through kriging (Hengl, 2009):

$$Y = X\beta + \lambda e + u \tag{4.2}$$

Here, a kriging model fits two parts concurrently, a regression $X\beta$ and a spatially correlated part λe based on a sample variogram function. The sample variogram function and its derivation will be explained later. The error term u, if there are no other omitted variables, should be just noise (Hengl, 2009). λe is a weighted average of detected spatial autocorrelation between and across sample locations, where the weights λ_i are derived from kriging weights (Montgomery & O'Sullivan, 2017; Shortridge, 2019):

$$\lambda e = \sum_{i=1}^{n} \lambda_i e_i = \begin{bmatrix} \lambda_1 & ... & \lambda_n \end{bmatrix} \begin{bmatrix} e_1 \\ . \\ . \\ . \\ e_2 \end{bmatrix} \tag{4.3}$$

These kriging weights λ_i are calculated based on semi-variances between the sample points, and between the sample points and the unsampled location (Shortridge, 2019). Semi-variances are measures of the squared residual differences between two observations, and are functions of distance intervals h. They capture proximity and similarity between observations (Bossong, 1999):

$$\gamma(h) = \frac{1}{2N} \sum_{i}^{n} \sum_{j:h_{ij}=h} (\epsilon_i - \epsilon_j)^2 \tag{4.4}$$

To compute the formerly explained weights, kriging requires a sample variogram function. A sample variogram function is a measure of spatial dependency in the sampled data (Wackernagel, 2013). To derive a sample variogram function, some intermediate steps must be taken.

The variogram, which measures the semi-variance between pairs of observations, must be constructed. Kriging uses the variogram to further generate a sample variogram, which is the averaged semi-variances for each distance interval h of the formerly generated variogram (Bivand et al., 2008). The dotted line in Figure 4.1 illustrates such a sample

variogram. Accordingly, three properties need to be estimated: the nugget effect, the sill, and the range. The nugget effect represents the discontinuity in semi-variance at the origin. If the sample variogram illustrates large semi-variances at small distances, then observations that are close to each other are displaying dissimilarities (Bossong, 1999). This suggests that a lot of variability in the dependent variable is not explained by the underlying predictors nor spatial autocorrelation.

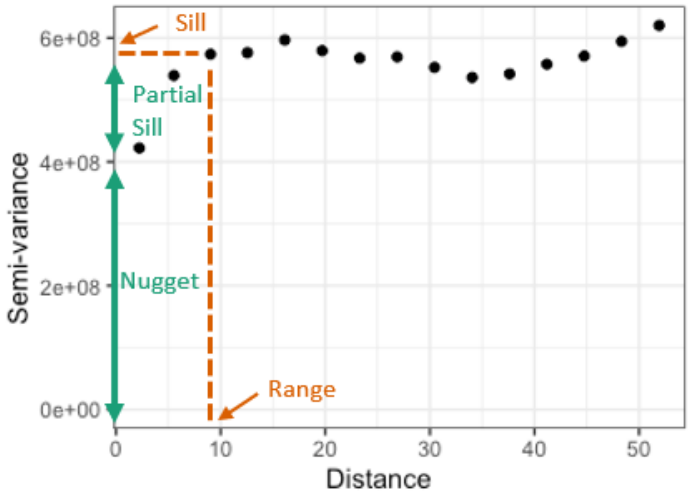

Figure 4.1. Example of a sample variogram, inspired by Guertal and Elkins (1996).

When the semi-variance in the sample variogram reaches a maximum, it is referred to as the sill. (Wackernagel, 2013). The sill net of the nugget effect is referred to as the partial sill (Cressie, 1988). From this peak point on, the sample variogram levels out. Lastly, the range is defined as the distance from which on the variogram reached its partial sill (Marko, Al-Amri, & Elfeki, 2014). As the semi-variance does not increase from this point on, a pair of points do no longer exhibit correlation beyond this distance (Webster & Oliver, 2007).

To derive an appropriate variogram function and thereby kriging weights, the abovementioned parameters need to be estimated in a model and fitted to the sample variogram (Bivand et al., 2008). While there are many available model types, most sample variograms display bounded semi-variances (Webster & Oliver, 2007). Therefore, it is common to choose the variogram model from a set of widely used bounded model types. The process of selecting the appropriate variogram model is presented in Chapter 5.4.3.

4.2.2 Random Forest

Random Forest (RF) is a popular supervised machine learning algorithm that operates by constructing an ensemble of decision trees. The algorithm is applicable to both regression and classification problems. To understand the advantages of tree ensembles, it is convenient to shed light on an essential weakness related to regular decision trees. While regular trees work great as a descriptive tool for the exact data used to create them, they generally suffer from low predictive performance compared to other statistical learning methods. This is mainly because regular decision trees suffer from high variance, meaning that the fitted trees may look substantially different, with only minor changes to the data (James, Witten, Hastie, & Tibshirani, 2013).

Before explaining how the RF algorithm deals with this, it is convenient to provide a brief description of how decision trees are built. More precisely, since this research mainly focused on regression problems, it was chosen to explain the process of constructing regression trees. The process of growing classification trees is, however, quite similar. Regression trees are built from the top down and consist of a series of splitting rules, referred to as internal nodes, that segment observations by dividing the predictor space. This involves creating J non-overlapping regions R_j by setting thresholds based on possible values for the different predictors. For each observation ending up in region R_j, the same prediction is given, which is the mean of the response variable for the observations belonging to that region. The goal is to select thresholds, thus divide the predictor space into regions, that minimize the *residual sum of squares* (RSS) (James et al., 2013):

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{4.5}$$

As previously indicated, the performance of decision trees may be substantially improved through different techniques for aggregating numerous trees. One such technique is *bootstrap aggregation* or *bagging*, a technique proposed by Breiman (1994). This is a technique employed by the RF algorithm, with the purpose of addressing the problem of high variance for regular decision trees. The technique starts by creating B bootstrapped datasets. These are datasets of the same size as the original, consisting of randomly selected samples S_b from the original data. The sampling of S_b is conducted with replacement, so

that the same observation may be picked more than once. Further, a decision tree f_b is fitted for each of the bootstrapped datasets S_b, creating a "forest" consisting of B decision trees. The predicted value for an observation x is obtained by averaging the predictions generated by the B trees, in the case of regression. For classification, the prediction is given by taking the majority vote.

$$y \leftarrow \hat{f}(x) \stackrel{\text{def}}{=} \frac{1}{B} \sum_{b=1}^{B} f_b(x) \tag{4.6}$$

Averaging the predictions for several trees yields a more robust consensus, which leads to the reduction in variance (Burkov, 2019; James et al., 2013).

In addition to bagging, the RF algorithm also includes a small tweak for decorrelating the decision trees. When constructing the trees, only a random subset of predictors are considered as candidates for the splitting criteria at the internal nodes. The one predictor of this subset that provides the greatest contribution to minimizing the prediction error is chosen as the predictor of consideration for the initial split, the root node. In the same way, a new random sample of predictors is conducted, and the best separator is chosen as the predictor of consideration for the next node, and so on. This random sampling forces the trees to look more different, by prohibiting the tree fitting process from using the strongest predictors in each bootstrapped dataset. This causes the predictions of the trees to be less correlated, and averaging many uncorrelated quantities yields a lower variance than averaging many highly correlated quantities (James et al., 2013). The number of predictors to randomly sample is a hyperparameter specified by the user prior to model fitting. This will be described further in Chapter 5.4.1.

The bootstrapped datasets, combined with the randomly sampled subsets of predictors, result in a wide variety of trees. This variety causes RF to suffer less from high variance problems than regular decision trees (James et al., 2013). The number of trees to grow is also a user-defined hyperparameter, which will be covered in Chapter 5.4.1.

Traditional decision trees are known for their simple interpretation, but the vast forest of trees resulting from the bagging procedure reduces the interpretability substantially (James et al., 2013). However, since bagging conducts sampling with replacement, there is a so-called *out-of-bag* sample for each bootstrapped dataset S_b, which are the observations

not included in S_b. This allows for calculating measures for *variable importance*. Variable importance is often assessed by either the *mean decrease in impurity* or by the so-called *permutation importance*. The former is a computationally fast measure of variable importance, but it has been shown to suffer from bias when predictors vary in their scale of measurement (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). The permutation importance is, for regression problems, assessed by recording each tree's prediction error on the out-of-bag sample. Then, the error is again recorded after iteratively permuting one by one predictor (Kuhn, 2019; Parr, Turgutlu, Csiszar, & Howard, 2018). Randomly permuting a predictor's values mimics its absence from the model, since its original relationship to the response variable is broken (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). The importance is measured as the increase in model error caused by permuting the predictor, compared to the baseline. These differences are averaged over all trees and normalized by the standard deviation, which yields the importance score (Kuhn, 2019; Parr et al., 2018). As will be presented in Chapter 5.3.1, the models of the first part of this study included location variables of different scales of measurement. This was considered a potential source of inflated importance measures in favor of the models with relatively higher spatial resolution. Because of this, the permutation importance was chosen as the measure for variable importance for this study. This was further motivated by Parr et al. (2018), who, in general, recommend this measure due to its robustness. Lastly, it should be mentioned that both methods may be subject to bias if the predictors are highly correlated since their associated importance gets spread over more than one predictor (Attanasi et al., 2020).

4.2.3 Geographical Random Forest

Geographical Random Forest (GRF) is a spatial analysis method, which is a variant of the famous RF algorithm. While the RF is a global and non-spatial algorithm not explicitly designed to address spatial heterogeneity, the GRF allows for this by fitting local sub-models for observations that are near in space. GRF is designed to bridge the gap between machine learning and geographical models, and is inspired by a method called Geographically Weighted Regression (GWR). However, unlike the GWR, the GRF is also capable of accounting for non-linear relationships between the response and the predictors along with modeling spatial autocorrelation (Kalogirou & Georganos, 2019a).

For constructing the GRF-model in R, the package *SpatialML* (Kalogirou & Georganos, 2019b), released in May 2019, was used.

Georganos et al. (2019) illustrate the difference between RF and the GRF by using a simple regression equation as a basis for the explanation.

$$y_i = \alpha x_i + \epsilon, \qquad i = 1, ..., n \tag{4.7}$$

Here, y_i is the observed value for observation i, αx_i is the predicted value by RF based on independent variables x, and e is the difference between the observed and predicted value. GRF alters the predictions generated by RF by weighing the location of observation i as represented by coordinates u_i and v_i (Georganos et al., 2019):

$$y_i = \alpha(u_i,\ v_i)x_i + \epsilon, \qquad i = 1, ..., n \tag{4.8}$$

Particularly, the concept of GRF involves fitting local models for different "neighborhoods" in addition to fitting a global model based on the entire training data. The neighborhood that the local model operates within is called a *kernel*, where the maximum distance between a data point and its kernel's border is called the *bandwidth*. The *SpatialML* package (Kalogirou & Georganos, 2019b) allows for defining two types of kernels; $fixed$ and *adaptive*. The former defines a fixed circle with a radius given by a user-defined bandwidth, while the latter defines the neighborhood flexibly by the n nearest neighbors, where n is given by the bandwidth. When the density of observations varies across space, using an adaptive kernel is advantageous (Georganos et al., 2019). This was the case for the wells in the dataset used for this study. Thus, at each location i, a local RF is fitted by only considering the n nearest observations in terms of space. These n observations are determined by the Euclidean distances between data points, referenced by their longitude and latitude coordinates. This leads to computing a RF for every single data point in the training set, where each model has its own performance (Georganos et al., 2019).

When using the GRF for predictions, the local and global models are fused by using a weight parameter ω. This allows for the extraction of local heterogeneity from the local models while merging it with the global model that is fitted on more data. Partly utilizing

the local model aims to lower bias, while partly utilizing the global model aims to lower variance. The weight is a user-defined hyperparameter, ranging from 1 to 0, where $\omega = 1$ leads to giving all the weight to the local model, and $\omega = 0$ leads to giving all the weight to the global model. When the GRF is used to predict production volumes on a new location, the nearest local model is used (Georganos et al., 2019).

For GRF to improve upon predictions from a regular RF, it is important to choose an appropriate bandwidth and weight parameter. As with the regular RF, the GRF also has hyperparameters for selecting the number of variables to randomly sample as candidates for each node split, as well as the number of trees to fit for each forest. The tuning process for selecting appropriate values for these hyperparameters will be outlined in Chapter 5.4.2. Lastly, the GRF also allows for computing the permutation importance.

4.3 Performance Measures

In this section, different metrics for quantifying model performance will be presented. Performance is assessed by measuring how well the predicted first-year production volumes match the actual volumes, for a set of observations held out from the model fitting.

4.3.1 Measures for Goodness-of-fit

For regression problems, one of the most common performance metrics is the root mean squared error (RMSE). The RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.9}$$

Here, y_i is the observed value and $\hat{y}_i$ is the predicted value, for the i^{th} observation. This metric will be smaller when predictions are closer to the true values. The RMSE expresses prediction errors in the same units as the response variable since the averaged squared differences are square rooted. This is beneficial for interpretability. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors compared to error metrics relying on absolute values. Because of this, RMSE is a well-suited metric when large errors are undesired (JJ, 2016). The RMSE is a scale-dependent

metric, which means it is not always clear what constitutes a good RMSE. An alternative performance metric is the mean absolute scaled error (MASE), which is given by:

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i - \bar{y}|} \tag{4.10}$$

Here, $\bar{y}$ represents the mean of the variable of interest. Thus, the residual of each prediction is scaled by a naïve baseline that predicts using only the in-sample mean. Thus, this metric provides an intuitive understanding of a model's predictive performance. If the MASE is less than one, the model serves as an improvement over using the in-sample mean to predict the response. The model generates perfect predictions if the MASE is zero. Conversely, if the MASE is greater than one, the naïve prediction outperforms the model (Hyndman, 2006).

4.3.2 Spatial Autocorrelation of Residuals

Since this research utilized spatial data, it was also of interest to measure how well the models managed to account for the spatial dependencies present. One of the most common measures of spatial autocorrelation is the cross-product statistic Moran's I. This statistic quantifies how related a variable's values are, based on their associated locations. Moran's I is calculated by:

$$I = (\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}})(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}) \tag{4.11}$$

Here, y_i is the ith observation of the variable of interest, $\bar{y}$ represents its mean, and w_{ij} represents the elements of a weight matrix W, with zeros on the diagonal. The weight matrix defines the location-based degree of connectedness between observations (Bivand et al., 2008). This weight matrix may be defined in several ways, and the choice will depend on the application. For instance, it may be defined by the travel time between two locations, a binary weight stating whether two locations are neighbors, or some measure of distance, like Manhattan- or Euclidean distance. Using the inverse distances conceptualizes that all cases are related to all other cases, but that the relatedness decays with increasing distances. Thus, the inverse Euclidean distance is appropriate for modeling

continuous data (Esri, n.d), such as geological properties, and was therefore chosen for this research.

Moran's I is represented by a scale ranging from -1 to $+1$, where values close to -1 represent highly dispersed observations (negative autocorrelation), and values close to $+1$ represent highly clustered observations (positive autocorrelation). Thus, the interpretation is similar to the regular correlation coefficient. When Moran's I is close to zero, there is little autocorrelation present (Dubé & Legros, 2014). In spatial modeling, the goal is to avoid spatial autocorrelation in the residuals of the fitted models. If one manages to include all relevant predictors of a spatial process that is described by a set of spatially autocorrelated predictors, the residuals will be random and not spatially dependent.

4.4 Model Development Framework

An important consideration in statistical modeling is to ensure a prudent framework for developing models that generalize well towards unseen data. To assess a model's ability to approximate a data generating process, or to obtain an understanding of how well it will perform in future applications, its prediction error must be computed on a set of observations held out from the model fitting process. Thus, it is the test error rather than the training error that is of interest, since the latter is usually an underestimation of the former (James et al., 2013). Most machine learning algorithms also require proper hyperparameter tuning to reach their full potential, with respect to performance. This process must also be conducted in a way which results in hyperparameters that are optimized for generalizability, rather than minimizing the training error (Burkov, 2019). To ensure this, a combination of the *validation set approach* and *k-fold cross-validation* (CV) was conducted for the first part of the research. For the second part, *nested CV* was conducted. These concepts will be explained in the remainder of this chapter.

The validation set approach involves splitting the data into a training- and test set. The training set is used for fitting models, and the fitted models are used to predict the response variable's values for the observations in the test set (James et al., 2013). Since the test set is just a held out fraction of the original dataset, each observation has an associated value on the response variable, which can be used for comparing its predicted and true value. This can further be used to compute performance metrics like RMSE and MASE. If one

were to compute performance metrics by comparing predicted and true values for the same set of observations that have already been used for model fitting, one would obtain overly optimistic estimates of model performance. This is because the models would, in a sense, memorize training examples and use this memory to "predict" the response (Burkov, 2019). A disadvantage with the validation set approach is that models are only trained on, for instance, half of the data. Since the model is fed fewer observations, the test error may be overestimated. In addition, the performance measurements may be heavily influenced by the specific split that constitutes the two sets. k-fold CV is a variant of the validation set approach that addresses these two issues. As with the validation set approach, k-fold CV also involves splitting the data. The difference is that instead of splitting it into a fixed train and test set, it splits the data into k folds of approximately equal size. The models are then trained on $\frac{k-1}{k} \times n$ observations, and the last held out fold is treated as a test set. This process is iterated k times, which leads to k models fitted on k different training sets, evaluated on k different test sets. At last, the performance is assessed by averaging the k CV-errors (James et al., 2013). Nested CV is merely a two-layer form of k-fold CV, where the CV loops are nested in order to ensure unbiased hyperparameter tuning (Raschka, 2018). How and why this was employed will be described in Chapter 5.2.

5 Model Development

This chapter of the thesis will describe the pre-processing of the data sets, considerations made when partitioning the data, as well as the process of configuring and tuning the different models.

5.1 Data Pre-processing

5.1.1 Preparing the Response Variable

When aggregating the data for the first 12 months of production per well, it was also chosen to keep and aggregate the variable *days_on_production*. This variable represented the number of days each well had been active during its first year of production. Keeping this in the dataset allowed for using the information to remove potential outliers. For instance, it was found that one of the wells had only been producing for 61 days during its

first year. At the same time, the first quartile value was 333 days, meaning that 75% of the wells had been producing for 333 days or more. Naturally, wells with extreme values in terms of production days also tended to exhibit substantially lower production volumes. Due to this, it was decided to apply the *IQR-rule* for omitting wells with extreme values in terms of production days. IQR stands for interquartile range and is a measure of how spread out a variable's values are. It is calculated as the difference between the third (Q3)- and first (Q1) quartile. The IQR-rule defines an observation as an outlier if its value is less than $Q1 - 1.5 \times IQR$ or greater than $Q3 + 1.5 \times IQR$. If this is the case, the observation is considered to deviate from the central values to an unreasonable extent (Chaudhary, 2019). The *days_on_production* variable has a natural upper bound of 365 days, and due to this, the IQR-rule was only applied to the lower values of the distribution. This led to removing 494 wells from the dataset since their number of production days during the first year was below the cut-off point of 303 days. Table 5.1 illustrates the difference in mean first-year production volumes, per year, between the two subsets created by the threshold value from the IQR-rule. As can be seen, the 494 wells below the cut-off point exhibited systematically lower production volumes. Removing wells with less than 303 days on production was considered a sensible measure for maintaining a clearer signal between the first-year production and the predictors.

Table 5.1

Mean first-year production volumes of the two subsets created by the IQR-rule.

	2011	2012	2013	2014	2015	2016	2017	2018
< 303	19,613	28,993	32,159	46,259	43,658	34,812	62,267	64,283
≥ 303	34,616	43,739	47,465	54,791	58,424	69,087	87,920	81,582

Additionally, the production volumes were scaled by dividing them by the fraction of each well's number of production days during the first year. This adjusted all production volumes to a volume as if the wells had been producing for 365 days.

5.1.2 Treatment of Outliers

Investigating the well-design data revealed some observations along the predictors *proppant* and $frac\_fluid$ that seemed unreasonably small. This is illustrated in Table 5.2.

Table 5.2

Quartiles of the proppant and $frac\_fluid$ variables.

	min	Q1	median	Q3	max
proppant	0	693.3	924.9	1099.9	9903.6
frac_fluid	0.455	17.280	20.286	27.597	177,512

By applying the IQR-rule to the lower values of the two variables, 8 and 6 observations were removed, respectively. The "extreme" values in the high end of the specter were considered more reasonable, since there was a more gradient transition from the third quartile to the maximum values, compared to the transition from the first quartile to the minimum values. Applying the IQR-rule to the high end of the specter would have led to omitting a total of 538 observations.

5.1.3 Treatment of Missing Values

In addition to a few outliers, the well-design data contained some missing values. The well-design variables of consideration for the first part of the analysis, *proppant*, $frac\_fluid$, and lat_length had 105, 335, and 2 missing values, respectively. Rather than omitting these observations from the data, it was chosen to impute the missing values. Several imputation techniques are available along a broad range of complexity and associated computational costs. For this study, a method called *bag imputation* was applied. This technique involves growing a bagged tree for each predictor by using all the remaining predictors in the dataset. When a sample with a missing value occurs, the bagged tree is used to predict its value. This is a powerful imputation technique with relatively high computational costs, but it was considered feasible due to the relatively low number of missing values in the dataset (Kuhn, 2019). The imputation model was constructed using only training data, in order to avoid information leakage between the training and test set. Further, since this technique utilizes information from all the other variables when imputing, it was chosen to keep the training data in its full dimension (except for some irrelevant metadata) until the imputation was conducted, even though most of these variables would not be used for the actual analysis later on. Outlier treatment was also performed prior to the treatment of missing values, in order to prevent the imputation

from being affected by outliers.

It should be noted that there is no guarantee that the imputation of missing values leads to improved model performance (Swalin, 2018). Due to this, it was also experimented with merely omitting the non-complete observations. It was, however, found that all models performed slightly better when bag imputation was conducted. As shown in Figure 3.2, the well-design variables exhibit a quite clear trend over time. It is reasonable to assume that this may have contributed to the missing values being imputed in a relatively representative manner when variables like *prod_start_y* were included in the imputation model.

5.2 Data Partitioning

As briefly mentioned in Chapter 4.1, working with spatial data requires special considerations concerning the data partitioning. Several studies have found that spatial autocorrelation may lead to considerably overoptimistic assessments of predictive performance, if data are split into training and test sets at random (Lovelace et al., 2019; Meyer et al., 2019). If spatial autocorrelation is present, it means that observations that are near each other in space tend to be more similar than distant observations. This breaks the assumption of independence between the training and test data because one will end up with observations in the test set that are very similar to observations in the training set, if the spatial proximity is not accounted for when splitting the data. This may cause information leakage between the training and test data (Lovelace et al., 2019). Thus, CV based on random data partitioning only validates a model's ability to reproduce the sampled data, and not its ability to generate predictions beyond it (Meyer et al., 2019). Several techniques for spatial CV have been proposed, aiming at obtaining more realistic performance assessments of spatial prediction models. These techniques involve splitting the data into spatially disjoint subsets, in order to avoid the information leakage caused by spatial autocorrelation (Lovelace et al., 2019; Meyer et al., 2019).

The first part of this research involved constructing different statistical models to obtain an understanding of which factors had been driving well productivity improvements over the last decade. Thus, it was more an investigative breakdown to understand what had been going on, where the models' ability to reproduce the data was of importance. Because of

this, CV was conducted with random data partitioning for this approach. The second part of the research involved fitting models for predicting production volumes at undrilled locations. Hence, this was more in the domain of predicting beyond the underlying data, and therefore, spatial CV was used for assessing the predictive performance. While the CV procedure itself is identical for both strategies, they differ in how the observations are split into different subsets. The following sections will present how this was incorporated when developing the models.

For the first part, the model fitting process began by implementing the validation set approach with a random 70/30-split of the data. Here, 70% of the observations represented the training set, while the remaining 30% represented the test set. Only the training set was considered through the process of model fitting, and the test set was only introduced when it was time to assess the performance of the final models. For the process of tuning hyperparameters, 5-fold CV was performed *within* the training set. Thus, the 70% of the data constituting the training set was further randomly split into five equally sized folds. Iteratively, the models were fit on four of the five folds, and their performance was evaluated on the last held out fold, the so-called validation set. Thus, the validation set was never used for fitting the models but served as an unbiased evaluation of model performance when considering different combinations of hyperparameter values (Shah, 2017). During the CV process, the models were trained iteratively with each specified combination of hyperparameters, and the associated CV-error was computed. In this way, the combination of hyperparameters that provided the best performance in terms of validation error across the five folds were identified. These values would represent the "optimal" combinations and were used for fitting the final models. The performance of the final models was at last assessed by being fed the matrix of predictors that constituted the test set, which allowed for comparing the predicted and actual values. More precisely, this performance was quantified by computing the MASE and RMSE. If the hyperparameters had been tuned with respect to the test set instead of the validation sets, they would have been unrealistically well adapted to the specific test set at hand. In a sense, the hyperparameters would be overfitted with respect to the test observations.

For the second part, a similar validation framework was used, but instead of partitioning the data at random, the proximity of observations was considered in order to ensure

spatially disjoint subsets of data. This was conducted by applying *k-means clustering* of the data points with respect to their location, referenced by longitude and latitude points. k-means clustering involves dividing a set of data points into clusters in a way that minimizes the total *within-cluster variation*. The total within-cluster variation is defined as the sum of squared Euclidean distances between each data point and its corresponding cluster centroid:

$$Minimize : \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{5.1}$$

Here, x_i is a data point belonging to cluster C_k, and μ_k represents the mean value of the points assigned to cluster C_k. This first involves choosing k, the number of clusters to define. Then, k data points are randomly chosen as the initial cluster centroids, and the data points are assigned to the cluster represented by the closest centroid in terms of Euclidean distance. For each of the clusters, a new centroid is computed by calculating the mean value of all the data points that constitute the cluster. This process is iterated until the centroids remain unchanged, or until the maximum number of iterations is reached (Jaiswal, 2018). For this analysis, the maximum number of iterations was set to 10, and k was set to 5. The resulting clusters are visualized in Figure 5.1.

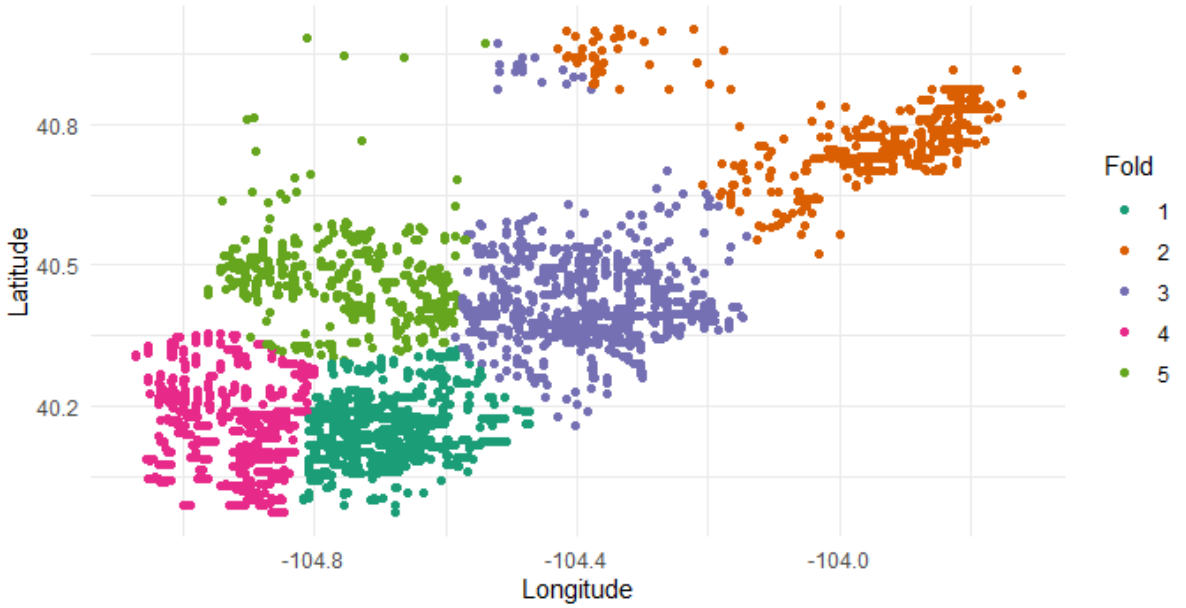

Figure 5.1. Spatially disjoint folds, generated by k-means clustering.

When fitting the model, one cluster was held out as test data, and the remaining four clusters were considered as training data. Since the different clusters exhibit different characteristics, the performance assessment would depend heavily on which of the five clusters were held out. Because of this, a validation set approach where one cluster would be the designated test set was considered too random, and it was instead chosen to apply 5-fold CV where each cluster represented a fold. Hence, for the first iteration, fold 1 in figure 5.1 would be held out as test data, and fold 2-5 would be used as training data, and so on. Since the clusters differed in size, the test set fraction varied from 13 to 26% of the data. In the same way as described earlier, 5-fold CV was again conducted within the training set. This is illustrated in Figure 5.2, where it can be seen that fold 1 from Figure 5.1 is held out. This two-layer form of CV is often referred to as nested CV (Raschka, 2018). The inner folds in Figure 5.2 were used for frequent unbiased evaluation of model performance when tuning hyperparameters. In other words, the purpose was identical to how 5-fold CV was used for hyperparameter tuning in the first part, just that the spatial proximity was accounted for. The hyperparameter combination found as most suitable from the "inner CV" was used for predicting on the held out fold.

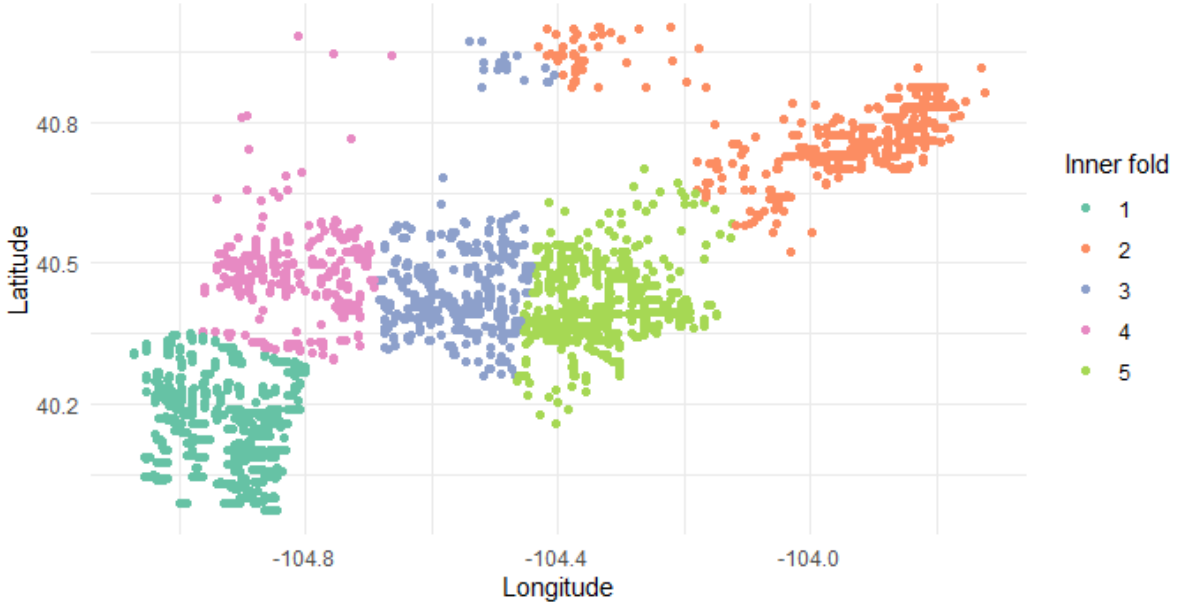

Figure 5.2. Spatially disjoint inner folds, generated by k-means clustering within fold 2 to 5 of Figure 5.1.

5.3 Model Configurations

Since the two parts of this study had different motives, their associated statistical models required custom configuration. This section will provide an overview of how the different models were specified.

5.3.1 Model Configurations - Part 1

The first part took a descriptive approach of modeling first-year production volumes, with the aim of obtaining an understanding of which factors had been driving well-productivity through the last decade. In other words, it used historical well-data to understand what had already taken place. For this purpose, well-design variables, alongside location as defined by pairs of longitude and latitude points, were used as explanatory variables. As established in Chapter 2, geolocation may serve as a suitable proxy for geological conditions, for descriptive purposes. The underlying dynamic here is that the location captures the tendency of similarly productive wells to be clustered, and their associated production volumes further indicate the quality of the geological conditions.

Inspired by Montgomery and O'Sullivan (2017), models of increasing spatial resolution were constructed in order to investigate the relative influence of well-design and location, on the first-year production volumes. On one end of the spectrum was the non-spatial model RF_ns. This was a RF consisting of only well-design variables, thus completely ignoring spatial dependencies. Next was the RF_fe, which used longitude and latitude points rounded to a precision of only one decimal. The reason for rounding the longitude and latitude points was to simulate less precise fixed effect regions, which assumed geological homogeneity within rectangles of approximately 11.1 $\times$ 8.5 kilometers (Veness, n.d.). This was done to make it distinct from the next model, the RF_xy. This model included longitude and latitude points as predictors with a precision of four and five decimals, respectively. This precision assumes geological homogeneity within rectangles of 1.1 $\times$ 8.5 meters. The purpose was to resemble a trend surface analysis, a form of linear regression where polynomials of spatial coordinates are used as predictors (Bivand et al., 2008). However, polynomials were not used since RF is capable of modeling nonlinearities (James et al., 2013). Figure 5.3 visualizes the difference in spatial resolution between RF_fe and RF_xy.

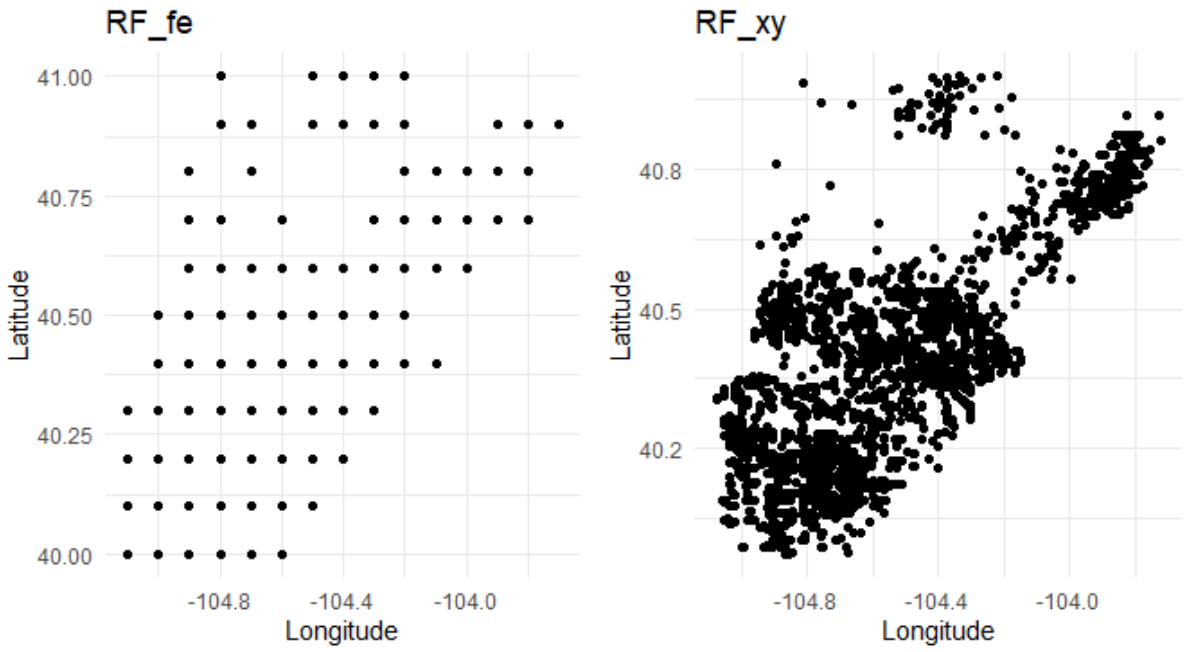

Figure 5.3. Illustration of the difference in spatial resolution between the RF_fe model (left) and the RF_xy model (right).

The next level of spatial sophistication was the *GRF*, which like *RF_xy* included the precise versions of longitude and latitude as predictors. In addition, the *GRF* used these longitude and latitude points as input for selecting the nearest observations when building the local sub-models. At last, since the abovementioned models were either somewhat experimental or relatively unestablished (GRF), the kriging model, *RK*, was included to serve as a benchmark for more established techniques. This method models spatial dependency by applying the matrix of kriging weights, as described in Chapter 4.2.1.

5.3.2 Model Configurations - Part 2

The second approach was of predictive nature, and involved predicting first-year production volumes at undrilled locations. Thus, it involved generating predictions beyond the area that the well-data was sampled from. As mentioned, longitude and latitude points capture the spatial clustering of wells. When the aim is to predict into undrilled acreage, this is, in fact, something that may end up only serving as noise because the geolocation itself is not directly linked to the conditions that are determinant of well-productivity. Since location serves as a proxy for geological conditions through the assumption that clusters of wells

indicate favorable geology, using it as a predictor for undrilled acreage would, to some extent, only suggest drilling close to already successful wells. Additionally, when using tree-based methods, some areas would unboundedly be highlighted as favorable due to the way the predictor space is divided by the node splits. These effects are easily illustrated by creating a heatmap based on predictions from a RF model with only longitude and latitude as predictors (figure 5.4).

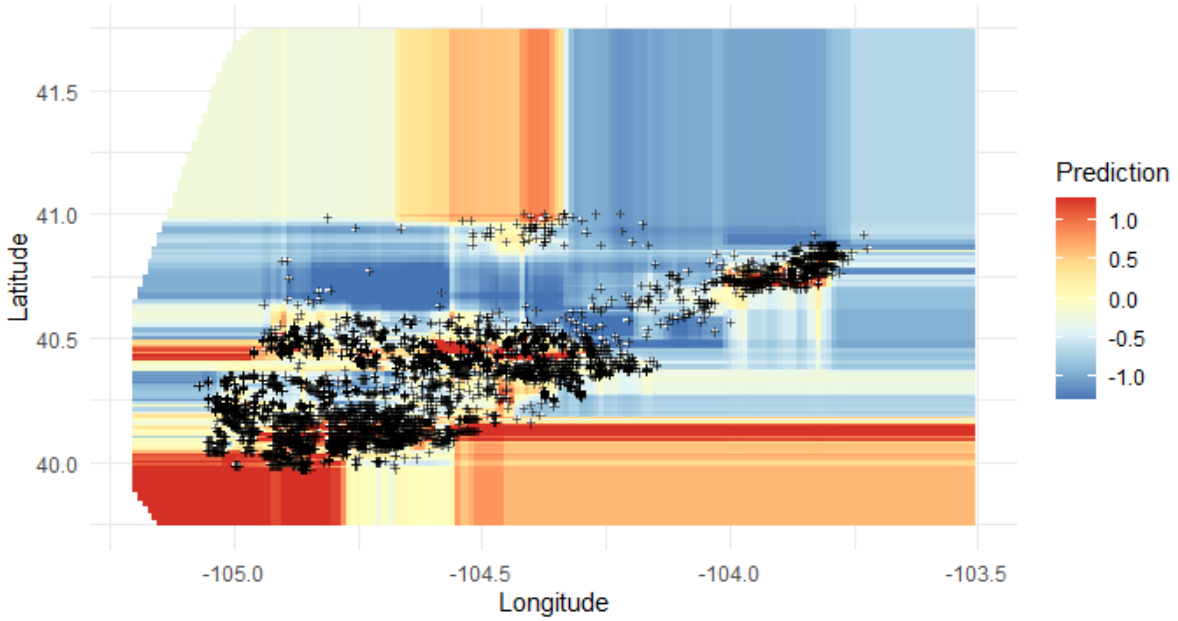

Figure 5.4. Heatmap generated by a RF with only longitude and latitude as predictors.

As may be seen from Figure 5.4, the patterns generated by using longitude and latitude as predictors seem quite unreasonable for generating predictions beyond the training observations represented by the black points. However, the patterns in between these points seem more reasonable. This is in line with what was mentioned in Chapter 2, that geolocation may be effective at reproducing the training data, but not for predicting beyond it (Meyer et al., 2019).

Due to this, predicting into undrilled acreage required spatial variables that were more directly determinant of the first-year production volumes. Thus, the geological variables presented in Chapter 3.3 were applied for constructing the models used for the second part of the study. Given the multitude of variables, and the fact that some of them represented quite similar properties, a framework proposed by Meyer et al. (2019) called *Spatial Forward Feature Selection* (Spatial FFS) was applied to systematically select relevant

predictors. This variant of feature selection is of particular interest for predictive purposes in a spatial spectrum, where spatial autocorrelation needs to be given careful consideration. Training a model on highly spatially autocorrelated predictors can put misleading emphasis on predictors that are insignificant for making spatial predictions beyond the training data. Spatial FFS addresses this by combining the technique of Forward Feature Selection and the previously elaborated spatial CV. Following a bottom-up approach, models are first trained on all possible pairs of potential predictors and validated through spatial CV. The best performing pair of variables are then used in an iterative process of improving spatially cross-validated model performance by further adding predictors. Due to the use of spatial CV rather than random CV, only variables that are relevant for predicting into spatially disjoint regions are selected. In addition, redundant predictors are also discarded, as with conventional techniques for feature selection (Meyer et al., 2019). Applying this technique resulted in selecting the variables: *isopach*, *slope_angle*, *topNio_surf*, *temp_gradient* and *topNio_msl*. To support the discussion in the previous paragraph, longitude and latitude were also included as candidates when conducting the Spatial FFS and were, as reported, not selected.

As mentioned in Chapter 2, well-design variables are well-established predictors of productivity, but they may not be sensible to include when the goal is to identify favorable geological settings at undrilled locations. This is because the well-design represents information that is simply unavailable until a well is drilled. This introduces an important implication because what was a favorable geological condition in 2011 was basically the same in 2018. However, a relatively good first-year production volume in 2011 was substantially lower than what was a relatively good production volume in 2018. This can be seen by the productivity trend visualized in Figure 3.1. As previously mentioned, numerous studies have shown that this increase in productivity may be attributed to the well-design variables. However, when these are not sensible to include as controls, the productivity trend must be accounted for in different ways in order to maintain a clearer signal between geology and production volumes. This was accounted for by grouping the production volumes by their associated year of initialization and standardizing them by subtracting the mean and dividing by the standard deviation of the associated year. To avoid information leakage between the training and test observations, the yearly mean and standard deviation were computed based on the training data. This correction aimed

to detrend the data by simulating that all wells were drilled in the same time period since the goal was to be better able to capture the quality of the geological resource rather than the quality of the well itself.

5.3.3 Summary of Model Configurations

Since the analyses of this research employed a multitude of model configurations, a summary of the different models is provided in Table 5.3 and 5.4.

Table 5.3

Model Configurations: Part 1.

Models	Y Variable	X Variables	Location Variables
RF_ns	*firstyear_prod*	*lat_length, frac_fluid, proppant*	N/A
RF_fe	*firstyear_prod*	*lat_length, frac_fluid, proppant*	long/lat, 1 dec [1]
RF_xy/ GRF/ RK	*firstyear_prod*	*lat_length, frac_fluid, proppant*	long/lat, 4/5 dec [2]

[1] Longitude and latitude in 1 decimal precision

[2] Longitude and latitude in 4 and 5 decimal precision

Table 5.4

Model Configurations: Part 2.

Models	Y Variable	X Variables	Location Variables
RF / GRF	*s_firstyear_prod* [3]	*isopach* *slope_angle* *topNio_surf* *temp_gradient* *topNio_msl*	N/A

[3] Standardized first year production volume

5.4 Model Tuning

This part of the thesis will describe the process of tuning the different models' hyperparameters. Hyperparameters can be thought of as "settings" that are manually specified by the data analyst prior to model fitting. It is essential to let the CV errors guide the hyperparameter tuning and not the test errors. If the test errors are used to decide upon the most suitable hyperparameter combination, the models may overfit on the test set. This leads to overly optimistic assessments of model performance, and hyperparameters tailored for the exact test set at hand rather than for generalizability (Burkov, 2019). The RK model does not directly have hyperparameters but requires selecting an appropriate variogram model, which also will be described in this section.

The RMSE was chosen as the metric to minimize when tuning hyperparameters because it, as mentioned in Chapter 4.3.1, is a suitable metric when large errors are undesired. That was the case for applications of the models in this study, as it reasonably can be argued that an error of 20,000 bbl is more than twice as bad as an error of 10,000 bbl.

5.4.1 Tuning the Random Forest

When using the *randomForest* package in R (Liaw & Wiener, 2002a), there are mainly four hyperparameters to consider when tuning a RF. These are *ntree*, *mtry*, *nodesize* and *maxnodes*. *ntree* represents the number of trees to grow. This should be set large enough to ensure that the error has converged and that every input sample has been predicted at least a few times. A large number of trees will not lead to overfitting, but it will increase training time. *mtry* represents the number of variables to randomly sample as candidates at each node split. The hyperparameter *nodesize* represents the minimum size of the terminal nodes of the trees. If this number is large, the fitted trees will be smaller and faster to train. At last, *maxnodes* represents the number of terminal nodes each tree can have. If this is not specified, trees can be grown as large as possible, subject to the limit given by *nodesize* (Liaw & Wiener, 2002b).

It is of common belief that one of the reasons for the RF algorithm's great popularity is that the default values for the abovementioned hyperparameters tend to empirically yield good predictive performance (Scornet, Erwan, 2017). Due to this, hyperparameter

tuning might not be the area that yields the highest returns in terms of performance versus time invested, for RF (Koehrsen, 2018). Because of this, the hyperparameters *nodesize* and *maxnodes* were kept constant at the default values of 5 and "unlimited", respectively. Further, it was decided to keep *ntree* constant at 500. This was decided after running tests to investigate the difference in errors at $ntree = 500$ and $ntree = 1200$, where it was found that the error was only marginally different. Thus, it was considered reasonable to conclude that the errors had converged at $ntree = 500$. The default value of *mtry* for regression problems, $\frac{p}{3}$, is often found to be optimal with respect to predictive performance (Koehrsen, 2018; Liaw & Wiener, 2002b). Here, p represents the number of predictors. However, Strobl et al. (2008) recommend considering different values of *mtry* when predictors are correlated. This was the case for the longitude and latitude variables of the first approach, where both pairs had a correlation of 0.74. For the second approach, $topNio\_msl$ and $topNio\_surf$ had a correlation of 0.92. Due to this, the range $mtry = [2 : p]$ was specified for each RF configuration. $mtry = 1$ was considered too random, and would potentially have required a substantial increase in *ntree* to stabilize the results. It was found that $mtry = 2$ yielded the lowest CV errors for each of the RFs in the first part, as well as for all the spatially disjoint folds for the second part's RF.

5.4.2 Tuning the Geographic Random Forest

As mentioned in Chapter 4.2.3, there are mainly four hyperparameters to consider when constructing a GRF; the bandwidth, *bw*, and the local weighting, *local.w*, in addition to *mtry* and *ntree* which is similar to the case of regular RF. The functioning of *bw* depends on which kernel type is specified. Since the density of the wells in this study varied through space, the adaptive kernel was chosen, as recommended by Georganos et al. (2019). In this case, *bw* represents the number of nearest neighboring observations used for building the local sub-models. *local.w* defines the weighting of the local and global models when generating the final predictions. Since fitting a GRF involves fitting local models for each observation in the training data, it is substantially more costly in terms of computational time than the regular RF (Georganos et al., 2019). Because of this, some measures were made in order to limit the number of possible hyperparameter combinations.

For the bandwidth, the range of possible values was specified as $bw = [25, 50, 75]$, for the

first approach. Initial experimentation revealed that bandwidths of 100 and above led to increased errors, as well as substantially increased training time. To exemplify, a GRF trained on 5000 observations with a bandwidth of 50 resulted in a model that occupied 1,35 gigabytes of memory in RStudio, while one with a bandwidth of 200 occupied 4,26 gigabytes. For the second approach, initial experimentation suggested that the model seemed to perform better with a larger bandwidth. Thus, the range was specified as $bw = [100, 150, 200]$. It is interesting to note that when spatial CV was conducted, larger bandwidths were preferred. This is most likely because spatial CV increases the distance between the training and test data, thus somewhat reducing the local signal's relevance compared to the case of random CV. Further, $mtry = [2, 3, 4]$ was specified for the model of the first part. These values were chosen since they were in the region of what was found as the "optimal" values for the regular RFs. It was chosen to limit the number of values compared to the tuning of the regular RFs, in order to decrease the computational demands. Since the models in the second part required an even more extensive validation process, $mtry$ was kept constant at 2. At last, $local.w = [1, 0.75, 0.50, 0.25]$, was specified as possible values for the local weight parameter. A value of 0 was considered unnecessary since this would result in increased tuning time while only building a global model equal to a regular RF. At last, $ntree$ was kept constant at 500, like with the regular RFs.

For the first part, models were trained and evaluated with each of these 36 hyperparameter combinations. This amounted to a run time of approximately 4 hours on a normal laptop. The best combination of hyperparameters was found to be $bw = 25$, $mtry = 3$, and $local.w = 0.5$. This means that the final model was constructed by building local sub-models based on the nearest 25 neighboring observations, that three predictors were randomly sampled as candidates for each split in the decision trees, and that the local and global model was weighted equally when fusing the final predictions.

For the second part, the outcomes of the validation process depended heavily on which subsets of the data were used for model fitting and hyperparameter tuning. Due to this, the hyperparameters selected by the tuning process are presented together with the performance metrics for the different spatially distinct subsets, in Table 6.4.

5.4.3 Preparing the Kriging Model

A variogram model should fit the sample variogram in a way that best represents the structure of the spatial autocorrelation in the data (Leroux, 2019). As mentioned in Chapter 4.2.1, there exists a set of widely used bounded model types. Some of the most common types are the so-called *Nugget, Bounded Linear, Circular, Spherical, Exponential, Gaussian*, and *Matérn* models (Bossong, 1999; Lichtenstern, 2013). These were all considered when "tuning" the RK model for this analysis. Technical details regarding these model types were considered beyond the scope of this work. The tuning process involved fitting variogram models to the sample variogram, cross-validating them individually, and comparing the RMSEs. In order to do this, the R packages *gstat* (Pebesma, 2004) and *automap* (Hiemstra, Pebesma, Twenhofel, & Heuvelink, 2008) were used. In addition to applying kriging to OLS residuals, it was also applied to RF residuals as an experiment. The best fit for the OLS residuals was achieved with the Spherical model, with a nugget of 340,731,464, a partial sill of 292,251,275, and a range of 7.75km. For the RF residuals, the pure Nugget model led to the best fit, with a nugget of 294,881,921.

6 Results

In this part of the thesis, the results from the final models of the two parts will be presented and briefly discussed.

6.1 Results - Part 1

After identifying the "optimal" hyperparameter combinations, as well as which variogram model to select for the RK, the models' generalizability with respect to unseen data was evaluated. The final models were fitted on the entire training set, with the hyperparameter values identified as yielding the best performance during the CV process. The models were evaluated by taking the matrix of predictors that constitute the held-out test set as input, in order to output predictions for first-year production volumes. The predicted values were then compared to the true values, and the performance metrics described in Chapter 4.3 were calculated. Table 6.1 summarizes the selected performance metrics for the different models.

Table 6.1

Summary of the different models' performance.

Model	RMSE	MASE	Moran's I
RF_ns	23,683.64	0.6567	0.2354
RF_fe	19,512.52	0.5284	0.1207
RF_xy	17,944.21	0.4794	0.0992
GRF	16,705.98	0.4451	0.0679
RK	18,853.69	0.5110	0.0481
RK_rf	18,500.60	0.4944	0.1045

Table 6.1 reveals that predictive performance improved as the spatial resolution of the RF models increased. The non-spatial RF_ns yielded the highest error metrics in terms of RMSE and MASE, and was also the model with the most spatially autocorrelated errors, revealed by Moran's I. All three of these metrics steadily decrease as the spatial sophistication of the models gradually increases through the RF_fe, RF_xy, and GRF. The GRF manages to account for spatial heterogeneity through fusing local sub-models with a global model. In this way, it is capable of extracting local signals (lower bias) while utilizing the global model's larger data basis (lower variance) (Georganos et al., 2019). As a result, the GRF's performance was superior to all of the other models. Interestingly, the well-established RK performed slightly worse than RF_xy in terms of RMSE and MASE, but the Moran's I signals that its residuals suffered from less spatial autocorrelation. This may imply that the RF_xy is better at modeling the relationship between the well-design variables and the response, but slightly worse at modeling the spatial processes. This suggests that there are two sources of bias at play, against which each of the models has its strengths and weaknesses. However, the gains of the former seem to outweigh the latter, resulting in a lower error for RF_xy.

Since the RF_xy performed better than RK in terms of RMSE and MASE, and since traditional RK is based on OLS residuals, it was experimented with constructing a kriging model based on residuals from RF_xy. The results from this model are summarized in the row corresponding to RK_rf, in Table 6.1. This model performed slightly better than the original RK in terms of RMSE and MASE, but with a higher Moran's I. However, its

performance was worse than the RF_xy. A potential reason may be that the initially fit RF_xy changed the spatial correlation structure of the residuals which potentially made the kriging ineffective. Indeed, the Moran's I indicates that more spatial autocorrelation is present in the residuals of RK_rf than for the traditional RK. To investigate this hypothesis, the sample variograms were compared. The left pane of Figure 6.1 illustrates the semi-variance of the OLS residuals. A partial sill is reached at a range of about 7.75 kilometers. Clearly, there seems to be autocorrelation that can be captured by the kriging model, up until distances equal to the detected range. The right pane reveals that the spatial autocorrelation structure was altered strongly in the case of RF_xy's residuals. With only a nugget effect and a range of zero, little obvious spatial autocorrelation can be detected. Accordingly, combining kriging with RF could not improve the performance, and since RK_rf was inferior to GRF and RF_xy, it was discarded. This was further motivated by the fact that the main reason for including RK was for it to act as a benchmark model for more traditional geostatistical techniques. The RK_rf would not serve as a sensible benchmark for this, due to its experimental nature.

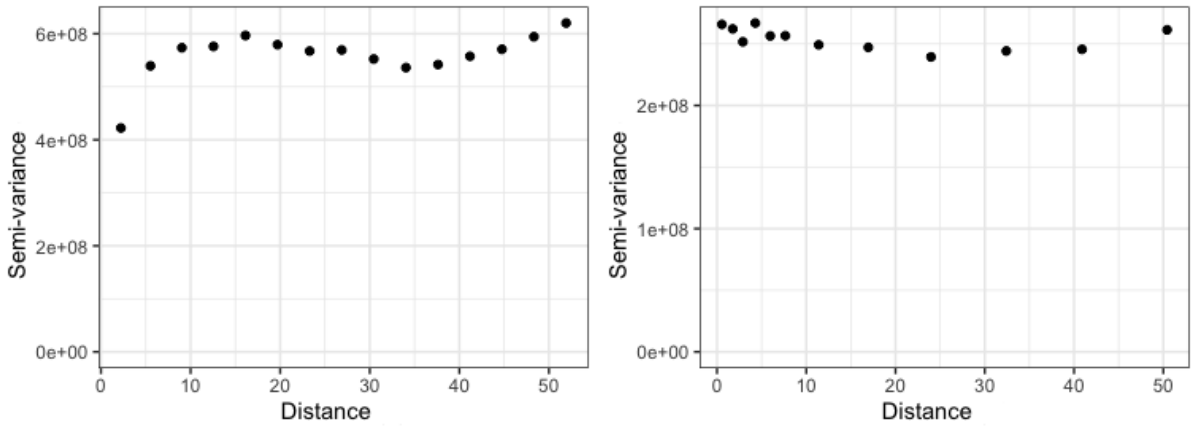

Figure 6.1. Sample variograms for RK and RK_rf (note that the y-axes differ). Left: Kriging of OLS residuals. Right: Kriging of RF_xy's residuals.

Further, a new manipulated test set was introduced to investigate the behavior of the models of different spatial resolution. This test set was generated by using forecasted well-design levels for 2020, and holding the well-design variables constant at these values. The forecasts were generated in a fairly simple manner, by fitting three linear regression models with each of the well-design variables averaged by year as the response, and the

year as predictor:

$$\widehat{well\_design} = \beta_0 + \beta_1 year \tag{6.1}$$

The models were fitted on yearly averages from 2011 to 2018. This yielded the forecasted values for 2020, summarized in Table 6.2.

Table 6.2

Forecasted average well-design levels for year 2020.

proppant [lbs/ft]	*lat_length* [ft]	*frac_fluid* [bbl/ft]
1210	8762	30

Replacing the well-design values in the original test set with these forecasted values created a scenario as if all these locations had been drilled with the predicted average 2020 well-design. Figure 6.2 visualizes the mean predicted production volumes for each of the models (note that the y-axis is truncated, beginning at 80,000 bbl). The figure reveals that the less spatially sophisticated models, on average, predicted substantially higher production volumes under these simulated circumstances. This suggests that these models generate more optimistic forecasts by giving the well-design variables more weight, compared to the models with higher spatial resolution. In other words, when more spatial variability is captured by the model, less weight is assigned to the well-design variables.

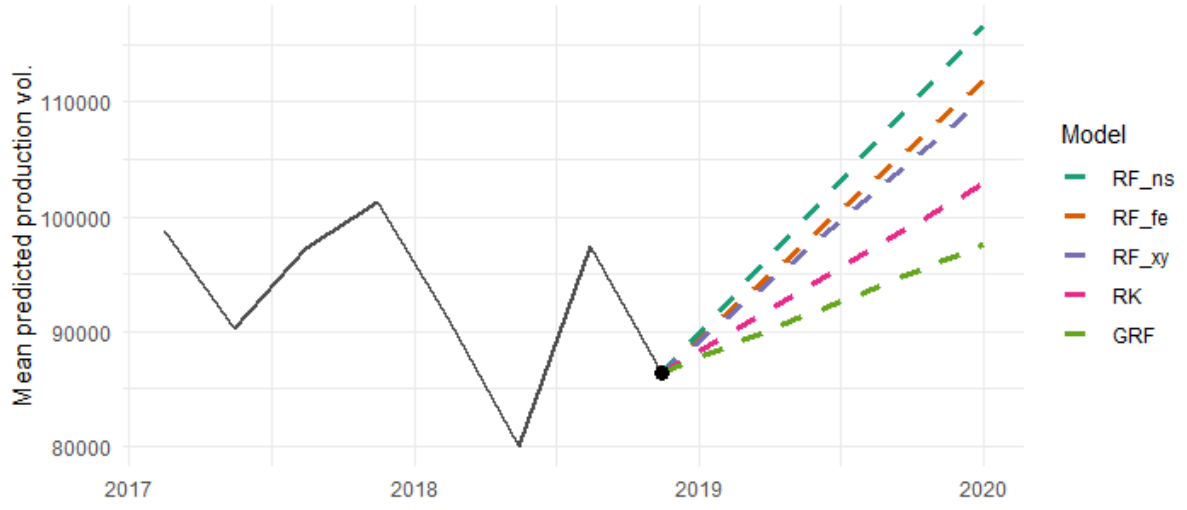

Figure 6.2. Predicted first-year production, using forecasted well-design levels for 2020.

To further investigate the different models' behavior, another manipulated test set was constructed. Here, the well-design variables were held constant at the average levels for the first quarter of 2011, for all observations in the test set. These average levels are summarized in Table 6.3.

Table 6.3

Average well-design levels of Q1-2011.

proppant [lbs/ft]	*lat_length* [ft]	*frac_fluid* [bbl/ft]
798	4089	24

The purpose of this was to isolate the impact of spatial effects from well-design, on well-productivity. Thus, this would illustrate the effect of high-grading practices. Figure 6.3 visualizes the mean predicted production volumes over time, when well-design is allowed to vary according to the data (green curves), and when held constant at Q1-2011 levels (orange curves), for the different models. The predictions are indexed to the first quarter of 2011.

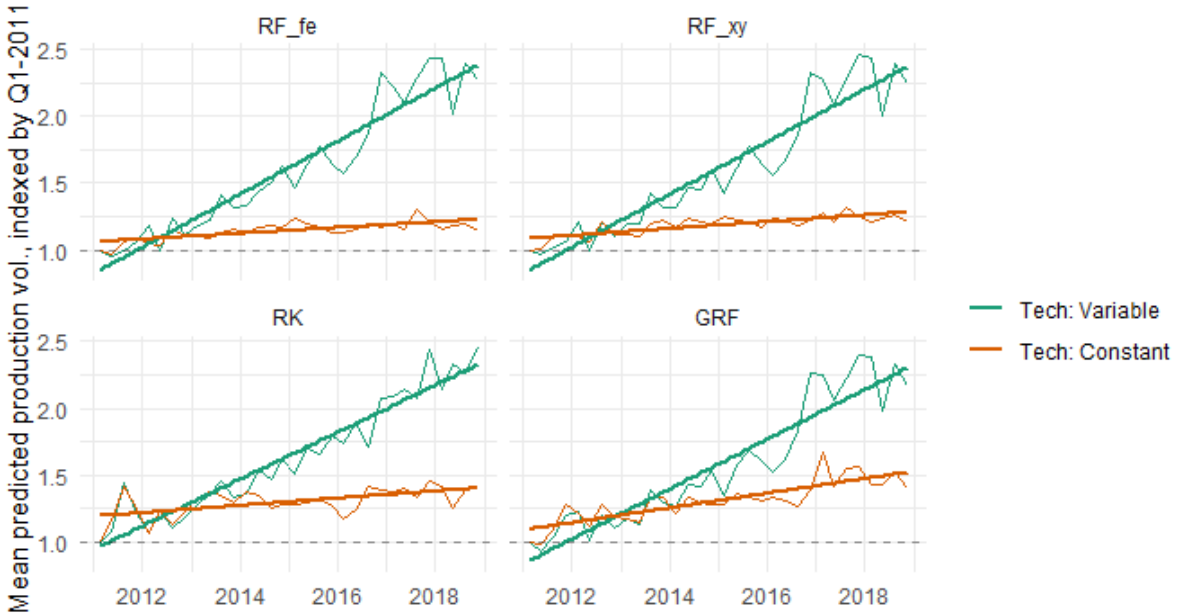

Figure 6.3. Comparison of predictions when well-design levels are held constant (orange curves), and when allowed to vary according to the data (green curves).

The figure reveals that, gradually, more of the improvements in terms of productivity are attributed to high-grading as the models become more spatially sophisticated. If one were to imagine a scenario where the government implemented regulations prohibiting operators from increasing well-design parameters beyond the levels summarized in Table 6.3, the orange curves would represent the different models' projections about the development in well-productivity. The models of lower spatial resolution present a far more pessimistic projection than the models of higher spatial resolution.

As a last way of investigating the behavior of the different models, it was chosen to compute their associated measures of variable importance. As outlined in Chapter 4.2.2, the permutation importance was chosen for this study, due to its general robustness. Strobl et al. (2008) mentions that interpretation of variable importance scores is only sensible when the number of trees, *ntree*, is set sufficiently large such that importance scores do not vary systematically with different random seeds. To ensure robust results, the importance scores and their associated mean and standard deviation were computed across ten different seeds. Figure 6.4 summarizes each predictor's average share of the total importance score, across the ten random seeds, for the different models. The standard deviations are presented underneath each of the fractions and imply robust results.

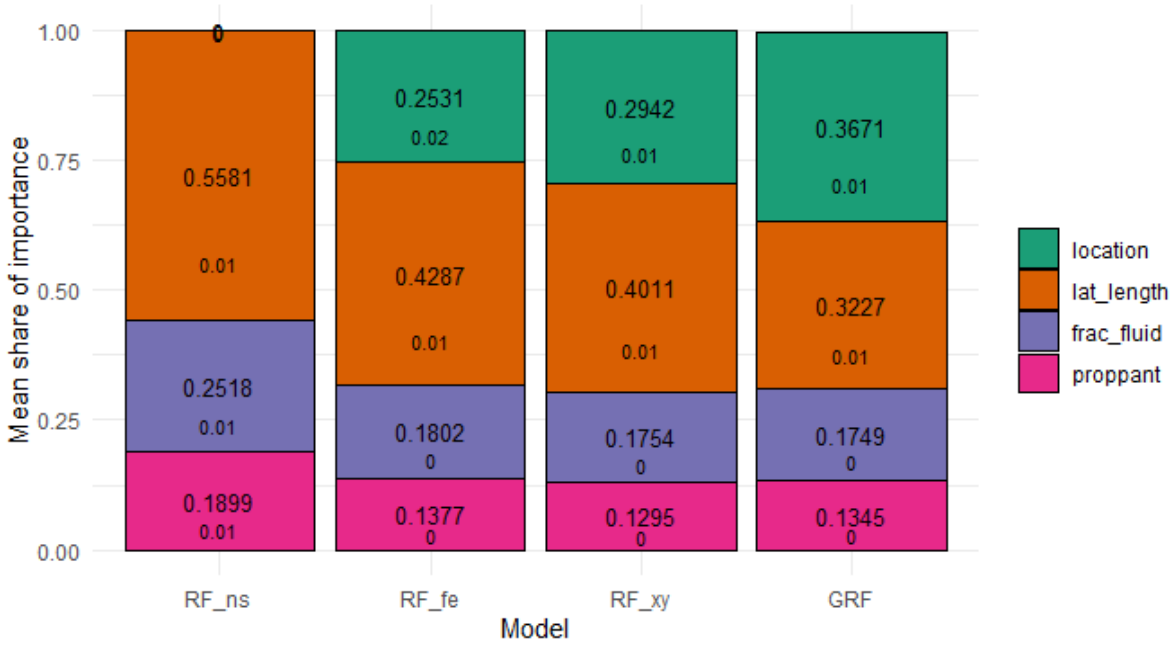

Figure 6.4. Mean relative importance of the variables in explaining the variation in production volumes.

As also mentioned in Chapter 4.2.2, importance measures may be subject to bias if predictors are highly correlated since their importance scores get spread over more than one predictor. In this case, the longitude and latitude variables correlated quite strongly, with a correlation of 0.74. Due to this, their importance scores were added together to what is represented as *location* in Figure 6.4. For the GRF, the fractions are calculated as the weighted sum of the importance scores from the global and local models. Since the models were weighted 50/50 when fusing the predictions, their importance scores were also weighted equally.

Figure 6.4 reveals how the importance of the different well-design variables gradually decreases as the models, more sophisticatedly, take location into account. Naturally, the non-spatial model may only attribute importance to the well-design variables, and as presented in Table 6.1, its associated performance is relatively poor. With increasing spatial resolution, more importance is attributed to location. Furthermore, this leads to improved performance across all reported metrics, as revealed in Table 6.1. Again, the results imply that the impact of well-design may easily be overestimated when the location is not, or inadequately, controlled for. This is especially true for the importance of lateral length, which is downgraded from an average share of importance of 0.56 to 0.32 throughout the spectrum of spatial resolution. Simultaneously, the importance of location increases from 0 to 0.37 through the spectrum of models. This implies that high-grading of geological conditions is an important factor in explaining the variations in well-productivity observed in the Niobrara through the last decade. It should be mentioned that "importance" in this case only implies that permuting a predictor's values leads to increased model error (Molnar, 2019), and does not directly imply causality.

The fact that location was attributed slightly short of 40% of the importance in explaining the variation in first-year production volumes suggests that a substantial share of the productivity gains from recent years are not directly transferable to new locations. This highlights the importance of selecting favorable locations for future drilling. This, and similar findings from previous research, motivated the second part of the analysis, which involved predicting the prospectivity of undrilled locations with the help of geological variables.

6.2 Results - Part 2

In this section, the results from the second approach will be presented and discussed. As previously mentioned, this approach involved entering the domain of predicting beyond observed data, thus requiring spatial CV to obtain realistic assessments of predictive performance (Meyer et al., 2019). Since the spatially disjoint subsets varied substantially in their characteristics, the performance assessment depended heavily on which subset was held out as test data. To obtain the best possible understanding of model performance, the performance metrics was therefore computed five times, once for each of the subsets as hold-out data. Table 6.4 summarizes the final models' performance across the five different test sets, as well as the hyperparameter values identified as "optimal" from the inner tuning process. For the GRF, *mtry* was kept constant at 2 in order to reduce computational time.

Table 6.4

Performance metrics from RF and GRF, across the 5 spatially disjoint folds.

RF			
fold	mtry	RMSE	MASE
1	2	0.9807	1.3146
2	2	1.3431	1.1572
3	2	1.2109	1.4391
4	2	1.1101	1.1860
5	2	1.0897	1.3322

GRF				
fold	bandwidth	local weight	RMSE	MASE
1	150	0.25	0.9657	1.3083
2	100	0.25	1.2527	1.0894
3	150	0.50	1.1340	1.3359
4	100	0.50	1.1237	1.3000
5	150	0.75	0.9812	1.2053

As can be seen from Table 6.4, the results from both models were rather disappointing. First of all, the fact that the MASE was greater than 1 for every region implies that one would be better off by naïvely predicting the mean first-year production volume in the available data for each region. Both the RMSE and MASE reveal that the predictive performance varied substantially between regions, but to a larger extent for RF than GRF. The GRF was also slightly better than RF, across all but one fold. As previously mentioned, the optimal bandwidths for GRF were substantially higher for this part, with an average of 130, than in the previous part where a bandwidth of 25 was chosen. This is sensible since the distances between the training and test data were bigger with spatial CV, which may have lowered the local sub-models' relevance. Further, the RMSE reveals that the error was highest for fold 2 and 3. Figure 5.1 reveals that these are the subsets of data that seem most dispersed. Moreover, the MASEs are of a substantially different magnitude than the MASEs computed in the previous part (see Table 6.1). This was expected, as the model assessment summarized in Table 6.4 mimics a scenario where first-year production volumes are predicted for more distant, undrilled locations. Since the locations are undrilled, it also means that well-design is simply non-available information. As identified in Chapter 6.1, these are variables that in total were attributed approximately 60% of the importance in explaining the variation in well-productivity. Not having that information available naturally makes the task more difficult. Since the RMSE is a scale-dependent metric, and the response variable of this part was standardized per year (as described in Chapter 5.3.2) to account for the productivity trend, the RMSEs reported in Table 6.4 are not directly comparable to the RMSEs from the previous part (Table 6.1).

Despite poor performance estimates, it was decided to plot heatmaps of the predictions to conduct a more qualitative assessment of the models' behavior. The final models used for generating the heatmaps were fitted with the hyperparameter values identified as best in the tuning process (see Table 6.4). For RF, *mtry* was set to 2 since it was chosen for all of the folds. For GRF, the average values across the five folds were specified, resulting in $bw = 130$ and $local.w = 0.45$. Figure 6.5 visualizes the generated heatmaps. Since both models generated a few very high predictions, the scale was narrowed to the range [-1, 1] in order to maintain a more informative visualization. Thus, predictions outside this range were downgraded to the nearest pole of the spectrum: either 1 or -1 standard

deviation from the mean of the yearly standardized production volumes.

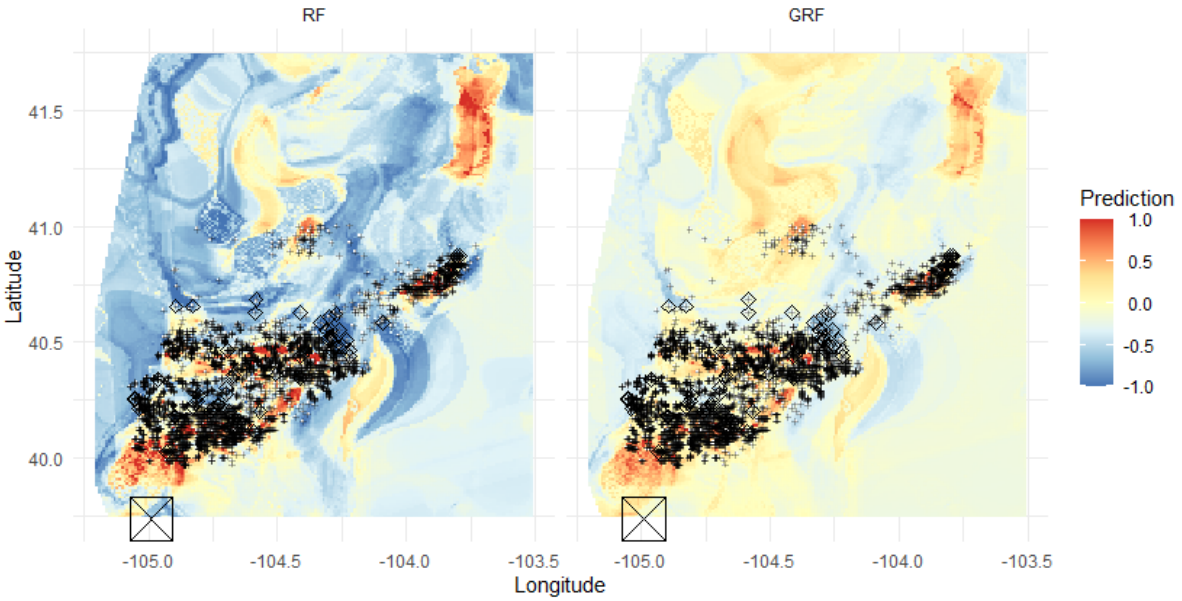

Figure 6.5. Heatmaps of predictions generated by RF (left) and GRF (right). The black points represent the training data. Points outlined with a diamond represent the least productive percentile of wells. The scale represents the yearly standardized first-year production of oil.

The heatmaps reveal several interesting aspects of model behavior. Firstly, the patterns suggest that there seems to be a quite strong consensus between the models. This is reasonable since the map to the left represents the global model, which is weighted 0.55 in the predictions that constitute the map to the right. However, the GRF tends to be more conservative in its predictions, which suggests that the GRF's local models seem to have a smoothing effect. Secondly, the patterns seem reliable as they change continuously, something one would expect a variable that is partly explained by geological conditions to do. Furthermore, it can be noticed that the coloring tends towards red in areas where many wells cluster, and that the lonely wells tend to be located in the more blue regions. To further investigate this, it was chosen to highlight the least productive percentile of wells by outlining them with a diamond icon. As can be seen from the figure, nearly all of these wells tend to lie in the less prospective regions of the maps. As a last qualitative assessment, it was chosen to further investigate some of the more prospective regions. In order to do this, the heatmaps were compared to the map presented in Figure 6.6. This

is a slightly modified version of a map presented by Han et al. (2019), which visualizes different well-known oil and gas fields in the Denver-Julesburg Basin. The modifications only involved removing information that was irrelevant for this study.

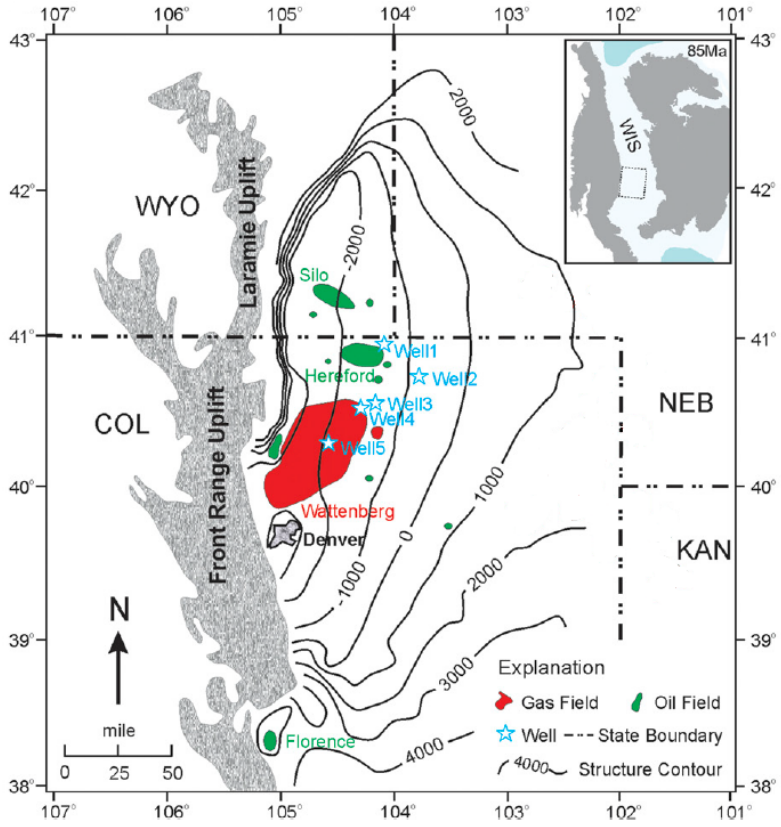

Figure 6.6. Map over oil and gas fields in the Denver-Julesburg Basin. (Han et al., 2019)

The square with the cross in the southwestern corner of the heatmaps in Figure 6.5 represents the approximate location of Denver City, Colorado. Just north of this square, a large, relatively red area can be identified. This region seems to align quite well with the large gas field Wattenberg, marked in red in Figure 6.6. Further, the S-shaped region tending towards orange/red at approximately -104.60, 41.25 in the heatmaps, and the small blob southeast of it seem to align quite well with the oil fields Silo and Hereford. These are the two green regions north of Wattenberg in Figure 6.6.

The abovementioned aspects of model behavior suggested that the models were onto something, but not with enough precision to yield reasonable results in terms of RMSE or MASE. Thus, instead of trying to pinpoint the exact first-year production volumes, it was decided to rephrase the problem as a classification task. The RF algorithm may as easily be applied to classification problems as regression problems by specifying a qualitative

response instead of a quantitative one. At the time of writing, the GRF algorithm was not yet developed for classification problems and modifying it was considered beyond the scope of this research. However, considering the relatively strong consensus between the RF and GRF, it was considered sufficient to carry out the following part with only RF. Hence, the response variable was transformed into a qualitative variable by categorizing the standardized first-year production volumes as either "low", "med" or "high". The quartiles of the standardized first-year production were chosen as somewhat arbitrary thresholds for conducting this categorization. This gave the thresholds summarized in Table 6.5.

Table 6.5

Threshold values for categorizing predictions.

Category	Threshold
low	$s\_firstyear\_prod < -0.6990$
med	$-0.6990 \leq s\_firstyear\_prod \leq 0.5085$
high	$s\_firstyear\_prod > 0.5085$

Categorizing predictions in this way resembles a technique commonly used in resource exploration called *Common Risk Segment Mapping*. This involves generating maps that use a traffic light color scheme to visually divide an area into different categories. In these maps, green regions represent areas with a high likelihood of success and warrants further exploration, yellow regions are uncertain, and red regions indicate that the play will probably not be viable, thus not consenting further exploration (Seg Wiki, 2020b). By employing a similar scheme, a prediction suggesting a standardized first-year production of 2 would be correctly classified as "high", even if the true label was 1,2. For early-phase applications like deciding on whether or not to allocate resources for further investigating an area, it is reasonable to assume that this may be a satisfactory level of precision. However, this error would have had a large negative effect on a performance metric like the RMSE and MASE.

Thus, the CV process was rerun with the task specified as a classification problem. This allowed for generating a *confusion matrix*, which further allowed for computing

the classification *accuracy* of the RF. A confusion matrix is a table that summarizes a classification model's prediction results on a held out set of observations, and the accuracy is a fraction given by the number of correctly classified samples divided by the total number of samples classified. One drawback of the accuracy metric is that it assumes that all error types are of equal importance (Burkov, 2019). Thus, it was decided to compute a performance metric that was chosen to be named the *severe error rate* (SER). The basis for this was that it was considered more severe to misclassify a well with an actual label of "low" as "high", and vice versa, than misclassifying either as "med". The severe error rate for "high" (SER_h) was computed by dividing the number of wells predicted as "high" that was actually "low", by the total number of wells predicted as "high". The SER for "low", SER_l, was computed contrariwise. The accuracy measures and the associated severe error rates across the different folds are summarized in Table 6.6.

Table 6.6

Accuracy measures of RF and GRF across the spatially disjoint folds.

Fold	Accuracy	SER_h	SER_l
1	0.3725	0.0652	0.2579
2	0.3815	0.1667	0.2713
3	0.4631	0.2230	0.1572
4	0.5289	0.2875	0.0352
5	0.3899	0.2171	0.1901

The accuracy measures in Table 6.6 reveal that the predictive performance still varied substantially across the different folds. At first glance, they still seem quite disappointing, especially for fold 1, 2, and 5. However, a naïve baseline model that classifies based on a random guess serves as a natural benchmark for measuring the performance. In a case like this, where the response variable has three possible categories, the expected accuracy would be $\frac{1}{3} = 0.333$ (Gauher, 2016). Thus, the classifier performed better than random guessing across all the folds. This may justify its use for at least certain applications. Since there was no reason to assume that the class distributions within the different folds would be known, neither a classifier performing a weighted guess based on the class distribution,

nor one always predicting the majority class were considered sensible baselines (Gauher, 2016). Further, the fact that the severe error rates never exceed 0.33 implies that the most severe error types are underrepresented compared to what would be the result of a random classifier. This implies that the predictions tended to lean in the correct direction of the low-med-high spectrum.

To get a look at the actual predictions behind the reported performance metrics, the underlying confusion matrices may be investigated. For simplicity, the matrices from the different folds were added together. This is presented in Table 6.7.

Table 6.7

Confusion matrix summarizing the true labels and the predictions generated by the RF classifier.

		Actual		
		low	med	high
Predicted	low	618	901	393
	med	767	1789	893
	high	135	349	234

The confusion matrix supports what was previously pointed out: the predictions tend to lean in the correct direction of categories. It also allows for computing the model's overall accuracy across the different folds, which is defined by dividing the sum of the diagonal by the sum of the matrix:

$$\frac{618 + 1789 + 234}{6079} = 0.4344 \tag{6.2}$$

Thus, overall, the model classifies better than chance by quite a good margin. This accuracy differs slightly from the mean of the "Accuracy" column in Table 6.6 (0.4272) since the number of observations in each fold vary because of the k-means clustering. It can also be seen that the "med" category impacts the accuracy quite heavily. One may, as inspired by the Common Risk Segment Mapping methodology, consider the "med" category as "uncertain" and ignore its contribution to the results. This allows for computing a "polarized" accuracy by only considering the corners of the confusion matrix

in Table 6.7. This yields an accuracy of:

$$\frac{618 + 234}{618 + 393 + 135 + 234} = 0.6174 \tag{6.3}$$

Again, this accuracy is by a good margin better than a random guess classifier, which in this case, would have an expected accuracy of $\frac{1}{2} = 0.5$.

With an established understanding of the RF classifier's performance, the model may be used to output a heatmap of its qualitative predictions. The resulting heatmap is presented in Figure 6.7. To investigate how the predictions aligned with the actual observed first-year production volumes, it was chosen to add a layer of points with the best (green) and poorest (red) five percentiles of wells on top of the underlying heatmap.

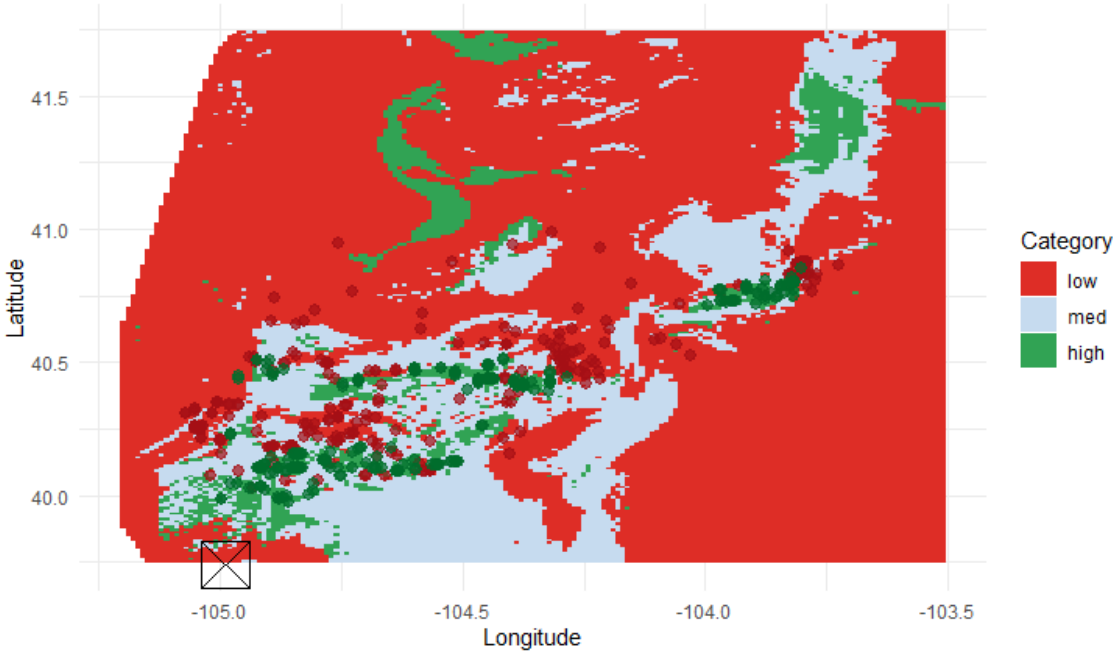

Figure 6.7. Qualitative heatmap of predictions from the RF classifier. The points represent the actual worst (red) and best (green) 5% of wells.

The heatmap indicates a quite strong consensus with the heatmaps presented in Figure 6.5. It can also be seen that the Wattenberg, Silo, and Hereford fields seem to be outlined, but in a more nuanced way than can be seen in Figure 6.6. Only small parts of the three fields are predicted as "high" when the thresholds are specified as presented in Table 6.5, but they seem to be clearly outlined inside of the borders defined by regions predicted

as "low". Furthermore, looking at how the predicted regions align with the top/bottom five percent of wells suggests that the model is successful at identifying the geological conditions that are associated with the more extreme production volumes. In a sense, this visualization serves as a graphical confusion matrix that further provides comfort regarding the model's reliability.

Lastly, the importance of using appropriate techniques for assessing the performance of spatial prediction models, through spatial CV, is a topic that has been stressed throughout this thesis. With this in mind, it was chosen to illustrate how one may easily be misled to accept a model that, at first glance, seems to have great predictive performance, due to being assessed by traditional CV with random data partitioning. Figure 6.8 visualizes the MASE and RMSE computed through traditional CV, relative to the metrics computed through spatial CV presented in Table 6.4. For the spatial CV, the x-axis represents which spatially disjoint subset of data is held out as the test set. For the random CV, the x-axis merely represents different random partitions of train/test sets.

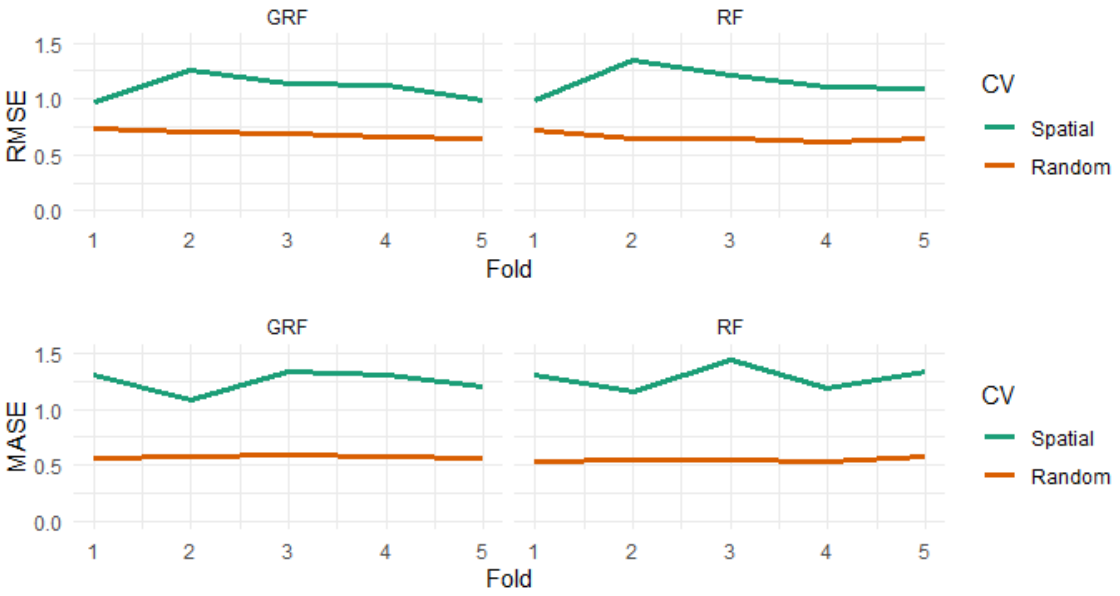

Figure 6.8. Performance metrics computed through random- and spatial CV. For spatial CV, the x-axis represents which spatially disjoint subset of data is held out as a test set. For random CV, the x-axis represents different random partitions of train/test sets.

It emerges from Figure 6.8 that the performance assessment of spatial prediction models may be drastically overoptimistic if conducted through traditional random CV. The RMSE

is considerably lower for each iteration, and the MASE suggests that the models are substantially better than naïvely predicting the mean production volumes in the available data. As previously mentioned, this happens because of information leakage caused by spatial autocorrelation, when the spatial proximity of observations is not accounted for when partitioning the data (Lovelace et al., 2019). It should be noted that the final heatmap would be similar, no matter the choice of CV procedure. Potentially, they would exhibit slight differences due to the different procedures for hyperparameter tuning, but the main point is that the final heatmap would be given substantially more confidence if the models were assessed with random- instead of spatial CV.

7 Discussion

This section will briefly summarize the main results and findings from each part of the research. Further, possible business applications of the insights and their associated managerial implications will be discussed. The findings of the descriptive first part will be discussed in light of similar previous research. Challenges related to the predictive and more experimental second part will be addressed, accompanied by discussions regarding how they could have been handled differently.

7.1 Discussion - Part 1

The first part of this research took a descriptive approach to investigate the relative impact of well-design and high-grading of geological conditions, on first-year oil production, by analyzing historical well-data. By employing statistical models of increasing spatial sophistication, it was found that model performance improved with the spatial resolution of the models. The performance measures also indicated that the more modern machine learning algorithms, Random Forest (RF) and Geographical Random Forest (GRF), performed slightly better than the well-established technique of Regression Kriging (RK) for this purpose. This is an interesting finding because these techniques are more data-driven, less restrictive, and require fewer assumptions than RK (Hengl et al., 2018). This result aligns well with findings from Schuetter et al. (2015) and Hengl et al. (2018), who also found that RF can obtain as good, or slightly better performance than kriging techniques. Additionally, with R packages such as *caret* (Kuhn, 2020), scripts for running

RF models may be coded very quickly, even for inexperienced R users. This improves the availability of analytical tools that work well for spatial data. The *caret* package is a set of functions designed for streamlining the process of creating machine learning models in R (Kuhn, 2019). Unfortunately, the GRF is not yet available in the *caret* interface.

Further, it was found that the relative impact of well-design variables may be substantially overestimated if models do not, or inadequately, account for spatial effects. This was assessed by investigating the models' behavior when applying them to different modified versions of the test data, inspired by Montgomery and O'Sullivan (2017), as well as by computing the associated variable importance of the various models. More precisely, it was chosen to compute the so-called permutation importance, which measures variable importance by observing the effect on model performance by randomly permuting each predictor. A phenomenon that is often referred to is the trade-off between predictive performance and model interpretability (James et al., 2013). At the same time, a model with weak predictive power may not be able to extract much useful information from the data, which further can lead to questionable conclusions about the impact of different variables (Breiman, 2001). It was considered that the RF and GRF, along with the permutation importance, seemed to balance these two considerations quite well. This was further motivated by Parr et al. (2018), who state that permutation importance is a reliable and broadly applicable technique. They recommend using permutation importance for all models because pitfalls related to interpreting regression coefficients, such as not normalizing input data and not properly handling correlated predictors, can largely be avoided.

There are several economic implications of the abovementioned findings. Firstly, more accurate models imply more reliable importance measures and interpretation of results (Breiman, 2001; Parr et al., 2018). As shown in Table 6.1, increasing the spatial resolution yielded models that generated more precise results. This provides more reliable variable importance measures, thus providing more reliable insight for future operations. It was also shown that the less spatially sophisticated models overestimated the impact of well-design variables. This highlights the importance of choosing appropriate analytical techniques for decision making. Operators who are continuously trying to improve their productivity may be misled into allocating their resources suboptimally if decisions regarding well-design

are based on, for instance, a non-spatial model like the RF_ns. A false impression of the returns from the upscaling of well-design variables may lead to increasing lateral lengths, or the volume of fracking fluid and proppants, to suboptimal levels where productivity improvements do not justify the additional costs. In fact, Luo et al. (2018) found that increasing the volume of proppants may actually lower production volumes under certain geological conditions. Given the shale oil industry's highly sensitive profit margins, optimizing the economics of extraction is of substantial importance for the operators' viability.

The same tendency of well-design variables' downgraded influence with an increasing spatial resolution of models was found by Montgomery and O'Sullivan (2017). In fact, with their kriging model, they found that the relative influence of location was even larger - slightly short of 50%. However, their analysis was based on well-data from the Williston Basin. Interestingly, Rystad Energy (2019) found that the influence of location was only 28%, but this was with a less spatially sophisticated model, a surface trend analysis (STA). This analysis was based on well-data from the Permian Basin. Montgomery and O'Sullivan (2017) also constructed a STA model, where the influence of location was about 25%, thus aligning well with the findings of Rystad Energy (2019). It might be reasonable to assume that Rystad Energy (2019) also may have attributed more influence to location if a model of higher spatial resolution had been applied, possibly so that the results from the two basins had aligned. Based on this, it may be hypothesized that the slightly higher influence of location in these areas is due to there being more available information than in the Niobrara. Thus, the high-grading efforts may be based on a more thorough foundation of knowledge and experience. Furthermore, it is interesting to note that the importance ranking of the well-design variables in this study (Figure 6.4) are in the opposite order of the results in the study conducted by Montgomery and O'Sullivan (2017). Their results imply that proppant is the most influential, and lateral length is the least influential. This may be due to the fact that it is not the same study area, or it may be because the variables exhibit nonlinear relationships that are better captured by the models employed in this study.

7.2 Discussion - Part 2

The second part of the study took a predictive approach to modeling first-year oil production volumes as a function of varying geological settings. At first, the task was formulated as a regression problem, which yielded quite disappointing results. After conducting a more qualitative assessment of the models' behavior through visually inspecting their generated heatmaps of predictions, it was hypothesized that a useful model could be obtained by lowering the precision and rephrasing the task as a classification problem. This resulted in a RF classification model that had an overall accuracy of 43,5% and performed better than chance (33,3%) for all of the five spatially disjoint subsets of data it was evaluated on. In addition, its rate of severe errors, defined as the rate of high (low) predictions that were actually low (high), was always, with a good margin, lower than what would be expected of a random guess classifier. This indicated that the errors tended to lean in the correct direction. At last, looking at how the predictions aligned with the most extreme wells in terms of productivity (Figure 6.7) suggested that the model was successful at identifying geological settings associated with such cases. These properties of the classifier may justify its use in certain business applications. For instance, mapping areas with a color scheme based on their prospectivity may be used as an initial step for constructing an exploration portfolio used for deciding which regions to allocate resources and further explore. For instance, using the color scheme from Figure 6.7, red areas may be discarded, while green and eventually, blue areas may be investigated further. Further investigation may involve investing in the acquisition of seismic data for the area of interest. Land-based acquisition of seismic data is characterized by being both difficult and very expensive, for instance, because it often involves negotiating with landowners for permits (Seg Wiki, 2020a). Being able to focus such investments towards the most prospective areas may help operators construct exploration portfolios cost-effectively. It should be mentioned that the thresholds specified for categorizing the response variable were chosen rather arbitrarily, in this study. In a real business setting, operators may set these thresholds based on their operational characteristics and risk preferences. For example, the lower bound for the "med" category could be specified so that the production volumes, with reasonable comfort, will be a break-even case at worst.

The results of the second part may, at least at first glance, not look too impressive. However,

several factors make predicting production volumes of unconventional wells, based on only geological variables, a tough task. Some of these were mentioned in Chapter 2, but the process of conducting this research has also unveiled some other challenging aspects. First of all, as was presented in Chapter 6.1, well-design variables may be attributed more than half of the importance in explaining the variation in well-productivity. This property of the data creates somewhat unclear signals between the geological setting and production levels, since there are several cases where the geological conditions are nearly identical, but with large spreads in the associated production volumes. As discussed earlier, this was the reason for grouping the data by year and standardizing it with respect to the yearly mean and standard deviation. This procedure improved predictive performance somewhat through mimicking a scenario where all wells were drilled in the same period, thus incorporating the trend in technological progress. However, there are still large intra-year differences in the well-designs that most likely explain a substantial part of the variation in the production volumes. This was not managed to account for in a sensible way, which obviously affected predictive performance. Experiments that involved training models with both geological and well-design variables were conducted, but it was considered to not completely fit the motive of this study's second part. It ended up creating a "location A is favorable given well-design X"-type of scenario, more than providing a robust overview of generally favorable geological settings. This approach required using imaginary well-designs for the undrilled locations because the well-design is information that is simply unavailable until a well has been drilled. For instance, appendix A1.1 illustrates two heatmaps generated by this approach, where one is based on the average well-design of 2015, and one is based on the forecasted well-designs presented in Table 6.2. As can be seen, the appearance of the heatmaps was, to a large extent, determined by these somewhat arbitrary chosen values for the well-design variables. This was not surprising given the relative influence of well-design identified in Chapter 6.1. Based on the abovementioned factors, it was decided to discard this approach. Additionally, it was found reasonable to assume that a real-life exploration workflow is characterized by first identifying favorable geological conditions, then deciding upon a suitable well-design, rather than the other way around.

Another factor that may make the task of the second part more challenging is the presence of bias in the dataset stemming from the different operators' competency and knowledge.

More precisely, operators conduct thorough assessments of the geological characteristics of regions prior to eventually drilling, for instance, by interpreting seismic and geophysical data (Wagener, 2018). This affects the data availability in that the production data are mostly available only at locations where the geological setting is relatively good. It was hypothesized that the classification model would be better at separating good and bad locations if a broader range of geological conditions had been represented in the data. Thus, based on Allen and Allen (2013), who present that it is unlikely for oil to be generated outside a temperature window of 60 to 130°C, it was decided to create imaginary first-year production volumes of zero bbl at locations in the geological data where the reservoir temperature was below 50°C. However, it was found that this only introduced noise and led to a lower predictive performance. It was also experimented with implementing less extreme variants of imaginary wells, in addition to different amounts of them, without successful results. At last, it should be mentioned that one of the steps conducted while pre-processing the response variable may have slightly amplified this bias. As explained in Chapter 5.1.1, all wells with less than 303 days on production during its first year were removed from the dataset. This might have led to omitting wells that, for instance, were plugged and abandoned after just a couple of weeks due to the operator realizing they will not be profitable. One potential reason for such a decision could be realizing that the geological setting is unfavorable, which in that case means the observation could have contained some useful information.

A last property of the data that may have made the task more challenging is the fact that the geographic coordinates of the bottomhole were used for connecting the geological variables to the production volumes. By doing this, one implicitly assumes that the geological conditions at the bottomhole's location are representative of the wellbore's entire lateral length. However, there are several occurrences of lateral lengths exceeding 10,000 feet ($\approx$ 3 kilometers) in the data. Given the geological variables' resolution of approximately 0.85 x 1.1km, it means that the geological variables are only representative for the last 500m of the wellbore, roughly. It is not trivial how to account for this properly, but a relatively simple solution that could have served as an improvement is to feature-engineer new location variables that are simply interpolations of the wellhead- and bottomhole locations.

7.3 General Discussion

The two parts of this research took quite different approaches to analyzing oil production in the Niobrara shale play. Given the different nature of the two approaches, special considerations had to be taken in terms of model configurations (Chapter 5.3) and data partitioning when assessing model performance (Chapter 5.2). Since the first part was a descriptive breakdown of historical well-data, it was considered sensible to use pairs of longitude and latitude as a proxy for geological conditions, as employed by Izadi et al. (2013); Montgomery and O'Sullivan (2017); Rystad Energy (2017a, 2017b, 2017c, 2019); Zhong et al. (2015). Further, since it did not involve generating predictions beyond the sampled data, it was considered adequate to conduct a traditional CV with random data partitioning. This was not the case for the second part, which included generating spatial predictions for new, distant, and undrilled acreage. In order to avoid overoptimistic estimates of predictive performance, spatial CV was conducted in line with recommendations from Lovelace et al. (2019) and Meyer et al. (2019). This was incorporated by using k-means clustering in order to account for spatial proximity of observations when partitioning the data, and then applying 5+5 fold nested CV. To illustrate the importance of spatial CV in such applications, it was shown that one easily might be misled to accept a model that falsely gives an impression of great predictive performance, due to being assessed by traditional CV (Figure 6.8). Such overoptimistic beliefs regarding a model's reliability may cause severe financial harm for operators if drilling decisions are based on its generated output. This may, for instance, involve underestimating the actual risk of a drilling prospect. Further, since the second part involved predicting into new acreage, it was, as discussed in Chapter 5.3.2, no longer purposeful to use location as a proxy for geological conditions. This was visualized in Figure 5.4. Figure 7.1 is an extension of Figure 5.4, which in a clear way illustrates the problem of including longitude and latitude in spatial prediction models.

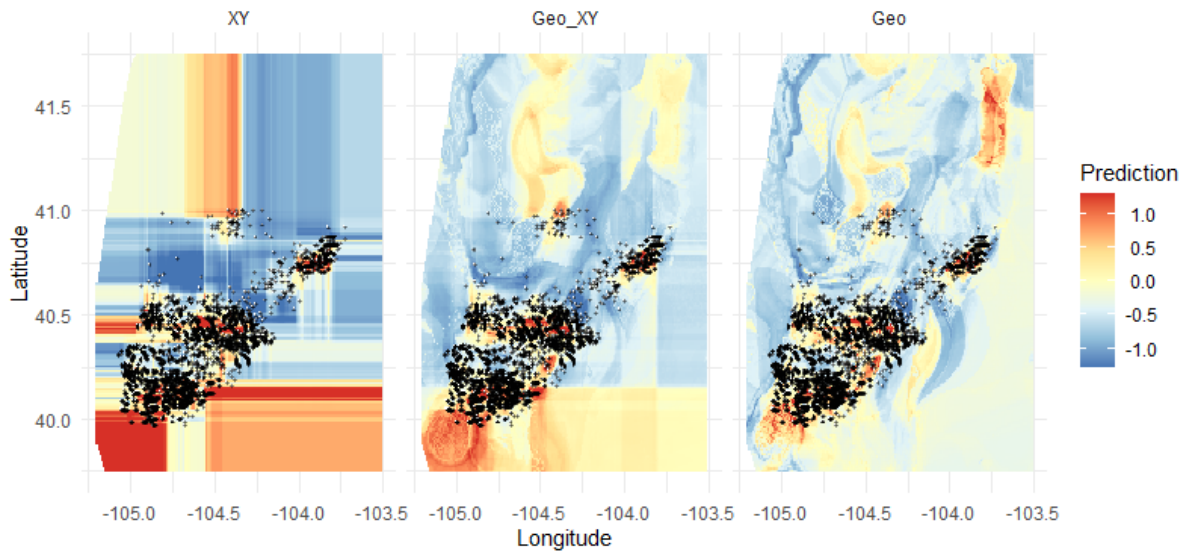

Figure 7.1. Heatmaps with the inclusion of different predictors. Left: Only geolocation variables. Center: Geolocation and geological variables. Right: Only geological variables.

The left-most heatmap was generated by a model using only longitude and latitude as predictors, while the right-most heatmap was generated by a model using only geological predictors. The heatmap in the middle used both longitude and latitude, as well as geological predictors. The middle figure reveals that unreasonable patterns emerge when longitude and latitude are included along with the geological predictors. These linear patterns are obvious artifacts and clearly align with the linear patterns in the left-most heatmap. This suggests that the geographical coordinates overrule the geological variables. In this study, the irrelevance of longitude and latitude, for generating spatial predictions, was identified by the spatial feature selection. However, since the artifacts often may be identified through inspecting the heatmaps, visual assessment of spatial predictions is stressed as a best practice by Meyer et al. (2019).

Furthermore, based on the relatively large importance of location identified in the first part, trying to identify those locations through predictive modeling was considered a natural next step in the workflow. The main reason for including the RK for the first part was for it to serve as a benchmark for more well-established techniques of spatial modeling. Since the results revealed that the RF and GRF performed slightly better than RK (in terms of RMSE and MASE), only those two models were employed for the second part. Additionally, since kriging is mainly an interpolative technique, extrapolation beyond the

observed data is not reliable (Christakos, 2012). Because of this, RK was not considered relevant for the second part's motive. Further, the GRF was no longer applicable when the task of part two was rephrased as a classification problem. However, the relevance of its speciality, namely its capability of modeling local signals by constructing local sub-models, may anyways be questioned for the second part. It is reasonable to assume that the local signals' relevance diminishes as the distance of the target locations increases. At some distance, it is even reasonable to assume that the local signals introduce more noise than information. The GRF was still slightly superior to the RF even when the performance was assessed through spatial CV, but this might not necessarily be the case if distances had increased further. Lastly, it should also be mentioned that the slight superiority of GRF came at a relatively high computational cost. For instance, training the final GRF used for generating the heatmap in Figure 6.5 took approximately 8 minutes, while generating the underlying predictions took about 35 minutes, on a standard laptop. For comparison, the same operations took 19 and 2 seconds with a standard RF, respectively.

7.4 Further Research

The results of the second part revealed that a classification model that, to a reasonable extent, outperformed random guessing could be achieved with help from the geological data at hand. As mentioned in Chapter 3.3, these variables were collected by digitizing maps from published studies. These were, however, only a handful of geological variables that may be determinant for oil production. For further research, it could be interesting to investigate whether the predictive accuracy could be improved further by including variables such as porosity and permeability. These, have as mentioned in Chapter 2, been identified as important variables for predicting production volumes. Further, the geological variables utilized in this study had only a spatial resolution of two decimals in their associated pairs of longitude and latitude points. It could also be interesting to see if predictive accuracy would improve with more precise data. This precision could potentially be further enhanced by using a location reference that is more representative for the entire wellbore, as discussed in Chapter 7.2. Further, as mentioned in Chapter 6.2, the conventional accuracy metric assumes that all error types are of equal importance. This was the reason for including the severe error rates when presenting the results of the RF classifier. However, for further research, it could be interesting to compute an

accuracy metric that incorporates the real-world impacts of the different error types in a more nuanced way. This could, for instance, be done by computing a *cost-sensitive accuracy* metric, as presented by Burkov (2019). This involves assigning a cost to the different types of misclassifications before calculating the accuracy. Computing such a metric would require a quite thorough assessment of the operational cost profile, in order to decide upon reasonable costs for the different errors. This was considered beyond the scope of this study.

The analyses conducted in this research have been of descriptive and predictive nature. An interesting extension of the findings in this research could be conducting a more prescriptive analysis based on the interplay of geological settings and well-design. Inspecting Appendix A1.1 reveals some interesting characteristics of this interplay. For instance, the area outlined by the green rectangles in Appendix A1.2 is found to be relatively good in both cases, while the center of the area outlined by the red rectangles is only relatively good with the 2015 well-design levels. This may suggest that the area within the green rectangle is characterized by conditions that can be made more fruitful by increasing the well-design variables. This area has a predicted first-year production volume of approximately 85,000 bbl in the 2015 scenario and approximately 125,000 bbl in the 2020 scenario. On the other hand, the area outlined by the red rectangles seems to be characterized by conditions where increasing well-design variables are not necessarily justified by improved production volumes. The center of this area has a predicted first-year production volume of about 90,000 bbl with the 2015 well-design while varying from about 80,000 to 100,000 bbl with the forecasted 2020 well-design. Further investigating this interplay could potentially unveil some interesting insights on how to optimize well-design under different geological settings, which may help operators improve their profits.

8 Conclusion

This research has focused on investigating oil production in the modestly studied Niobrara shale play. Two different analyses were conducted using three sources of data (production data, well-design data, and geological data). The first part of the research employed machine learning techniques to conduct a descriptive analysis of historical well-data, aiming at identifying drivers of well-productivity. Models of increasing spatial resolution were

employed in order to isolate the effect of high-grading of geological conditions from well-design choices. The models employed were different Random Forest (RF) configurations, a RF variant specifically designed for spatial data called Geographic Random Forest (GRF), and the widely recognized technique of Regression Kriging (RK). The spatial effects were in this first part modeled by using geolocation variables (pairs of longitude and latitude) as a proxy for geological conditions. Firstly, it was found that the higher the spatial sophistication of the models, the greater the model performance. The best performing model was the GRF, followed by the RF configuration of highest spatial resolution. It is interesting to note that the more data-driven RF and GRF outperformed the well-established RK model. Further, it was found that models that do not, or inadequately, account for spatial effects attribute too much importance to well-design variables. This was illustrated in several ways: first, by applying the models to different modified versions of the test data, then by computing measures for variable importance. The variable importance computed by the most spatially sophisticated model, the GRF, revealed that location was attributed slightly short of 40% of the total importance. Thus, the results implied that high-grading of geological conditions is a factor of substantial importance in explaining variations in the first-year production volumes of oil. This is crucial insight for operators who are aiming at improving their productivity. If decisions regarding well-design are based on models that fail to incorporate spatial effects, operators may end up increasing well-design variables to sub-optimal levels.

Given the substantial importance of high-grading identified in the first part of the research, the second part took a more predictive approach aiming at identifying undrilled locations with favorable geology. This part of the analysis used production data along with geological data, in order to model first-year production volumes as a function of geological conditions. For this, RF and GRF were employed. It was found that the results were somewhat disappointing when the task was formulated as a regression problem. This supports findings from previous research suggesting that it is difficult to model well-productivity as a function of geological properties only. However, instead of trying to pinpoint the exact first-year production volumes, the precision was lowered by rephrasing the task as a three-class classification problem. Thus, locations were predicted as either "low", "medium", or "high" in terms of first-year production volumes. This resulted in a RF classifier with an overall accuracy of approximately 43%, thus outperforming random guessing to some

extent. Further, an error metric that was chosen to be named the "severe error rate" revealed that the errors of the model tended to lean in the correct direction. A visual assessment of the model's generated heatmap of predictions suggested that the model was highly capable of identifying geological settings associated with the most extreme wells. These properties of the classifier may justify its use in certain business applications. For instance, it may be employed as a tool in the initial steps of constructing an exploration portfolio by filtering out interesting areas to invest in further investigation of. Since this part of the research involved generating spatial predictions beyond the observed data, a special technique for assessment of spatial prediction models, spatial cross-validation (CV), was employed. The importance of this was illustrated by computing performance metrics through both traditional and spatial CV, showing how the performance assessment may be substantially overoptimistic if conducted wrongly.

It is in the authors' belief that this research has been conducted in a transparent way, where necessary measures for ensuring the validity of results have been taken along the way. We believe that this research has contributed to providing more insight into the Niobrara shale play, and unconventional oil production in general. Additionally, it contributes to the literature by employing the GRF to a new domain. In fact, it contributes to assessing the applicability of GRF in general, which, according to our knowledge, is only employed in three published papers. In these studies, it is applied for modeling population density and human mobility. Further, to our knowledge, this is also the first application of spatial CV in the domain of unconventional oil production. At last, it supports findings from other studies that have found machine learning techniques to yield as good, or slightly better performance than the widely recognized kriging techniques.

References

Ahmed, U., & Meehan, D. N. (2016). *Unconventional oil and gas resources.* CRC Press.

Allen, P. A., & Allen, J. R. (2013). Basin analysis: Principles and application to petroleum play assessment. Wiley-Blackwell.

Amaechi, U., Ikpeka, P., Xianlin, M., & Ugwu, J. (2019). Application of machine learning models in predicting initial gas production rate from tight gas reservoirs. *Rudarsko-geološko-naftni zbornik*, *34*, 29-40. doi: 10.17794/rgn.2019.3.4

Attanasi, E. D., Freeman, P. A., & Coburn, T. C. (2020). Well predictive performance of play-wide and subarea random forest models for Bakken productivity. *Journal of Petroleum Science and Engineering*, *191*, 107150. doi: 10.1016/j.petrol.2020.107150

Bhattacharya, S., & Mishra, S. (2018). Applications of machine learning for facies and fracture prediction using bayesian network theory and random forest: Case studies from the Appalachian Basin, USA. *Journal of Petroleum Science and Engineering*, *170*, 1005–1017. doi: 10.1016/j.petrol.2018.06.075

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2008). *Applied spatial data analysis with R.* Springer.

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R.* Springer. doi: 10.1007/978-1-4614-7618-4

Bossong, C. (1999). *Overview and technical and practical aspects for use of geostatistics in hazardous-, toxic-, and radioactive-waste-site investigations* (Vol. 98) (No. 4145). US Department of the Interior, US Geological Survey. Retrieved from `https://pubs.usgs.gov/wri/1998/4145/report.pdf`

Breiman, L. (1994). Bagging predictors. Retrieved from `http://nlp.postech.ac.kr/~project/Course/CS704/related_papers/BAGGING_PREDICTORS.pdf`

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199-231. doi: http://dx.doi.org/10.1214/ss/1009213726

Burkov, A. (2019). *The hundred-page machine learning book.* Andriy Burkov.

Chaudhary, S. (2019). *Why "1.5" in IQR method of outlier detection?* Retrieved from `https://medium.com/mytake/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097`

Christakos, G. (2012). *Modern spatiotemporal geostatistics.* Dover Publications. Retrieved from `https://books.google.no/books?id=R5usceskjCEC`

Clarkson, C., Jensen, J., & Chipperfield, S. (2012). Unconventional gas reservoir evaluation: what do we have to consider? *Journal of Natural Gas Science and Engineering*, *8*, 9–33. doi: 10.1016/j.jngse.2012.01.001

Collins, R., & Adams-Heard, R. (2019). *Flaring, or why so much gas is going up in flames.* Retrieved from `https://www.bloomberg.com/news/articles/2019-08-30/flaring-or-why-so-much-gas-is-going-up-in-flames-quicktake`

Covert, T. (2015). Experiential and social learning in firms: the case of hydraulic fracturing in the Bakken shale.
doi: 10.2139/ssrn.2481321

Cressie, N. (1988). Spatial prediction and ordinary kriging. *Mathematical geology*, *20*(4), 405–421.

Curtis, T., & Montalbano, B. (2017). Completion design changes and the impact on US shale well productivity. *Oxford Institute for Energy Studies: Oxford, UK*. doi: 10.26889/ei212017

DiLallo, M. (2018). *The 5 companies dominating the Niobrara shale*

play. Retrieved from https://www.fool.com/investing/2016/08/25/the-5-companies-dominating-the-niobrara-shale-play.aspx

DiSavino, S. (2019). *Explainer: Why are U.S. natural gas prices in Texas below zero?* Retrieved from https://www.reuters.com/article/us-natgas-pipelines-flaring-explainer/explainer-why-are-us-natural-gas-prices-in-texas-below-zero-idUSKCN1RL2NL

Dong, Z., Holditch, S., & McVay, D. (2013). Resource evaluation for shale gas reservoirs. *SPE Economics & Management*, *5*(01), 5–16. doi: 10.2118/152066-MS

Dubé, J., & Legros, D. (2014). Spatial autocorrelation. In *Spatial econometrics using microdata* (p. 59-91). John Wiley Sons, Ltd. doi: 10.1002/9781119008651.ch3

EIA. (2019). *U.S. crude oil and natural gas proved reserves, year-end 2018.* Retrieved from https://www.eia.gov/naturalgas/crudeoilreserves/pdf/usreserves.pdf

EIA. (2020a). *U.S. field production of crude oil.* U.S. Energy Information Administration. Retrieved from https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=MCRFPUS2&f=M

EIA. (2020b). *U.S. natural gas marketed production.* U.S. Energy Information Administration. Retrieved from https://www.eia.gov/dnav/ng/hist/n9050us2a.htm

Elliot, R., & Santiago, L. (2019). *A decade in which fracking rocked the oil world.* Retrieved from https://www.wsj.com/articles/a-decade-in-which-fracking-rocked-the-oil-world-11576630807

Esmaili, S., & Mohaghegh, S. D. (2016). Full field reservoir modeling of shale assets using advanced data-driven analytics. *Geoscience Frontiers*, *7*(1), 11–20. Retrieved from 10.1016/j.gsf.2014.12.006

Esri. (n.d). *Modeling spatial relationships.* Retrieved from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/modeling-spatial-relationships.htm#GUID-729B3B01-6911-41E9-AA99-8A4CF74EEE27

Fu, D. (2019). Guest editorial: Unlocking unconventional reservoirs with data analytics, machine learning, and artificial intelligence. *Journal of Petroleum Technology*, *71*(01), 14–15. Retrieved from 10.2118/0119-0014-JPT

Gauher, S. (2016). *Is your classification model making lucky guesses?* Retrieved from https://blog.revolutionanalytics.com/2016/03/classification-models.html

GEBCO. (2019). *GEBCO_ 2019 grid.* Retrieved from https://www.gebco.net/data_and_products/gridded_bathymetry_data/gebco_2019/gebco_2019_info.html

Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., ... Kalogirou, S. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, *0*(0), 1-16. doi: 10.1080/10106049.2019.1595177

Gold, R. (2014). *Fracking gives U.S. energy boom plenty of room to run.* Retrieved from https://www.wsj.com/articles/fracking-gives-u-s-energy-boom-plenty-of-room-to-run-1410728682

Goovaerts, P. (2006). Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point poisson kriging. *International Journal of Health Geographics*, *5*(1), 52.

Government of British Columbia. (n.d.). *Conventional versus unconventional oil and gas.* Retrieved from https://www2.gov.bc.ca/assets/gov/farming

-natural-resources-and-industry/natural-gas-oil/petroleum-geoscience/conventional_versus_unconventional_oil_and_gas.pdf

Guertal, E., & Elkins, C. (1996). Spatial variability of photosynthetically active radiation in a greenhouse. *Journal of the American Society for Horticultural Science*, *121*. doi: 10.21273/JASHS.121.2.321

Han, Y., Horsfield, B., Mahlstedt, N., Wirth, R., Curry, D. J., & LaReau, H. (2019). Factors controlling source and reservoir characteristics in the Niobrara shale oil system, Denver Basin. *AAPG Bulletin*, *103*(9), 2045-2072. doi: 10.1306/0121191619717287

Hart Energy. (2011). *The D-J Basin's Niobrara vs. the Bakken and Eagle Ford—how the plays compare.* Retrieved from https://www.hartenergy.com/opinions/d-j-basins-niobrara-vs-bakken-and-eagle-ford-how-plays-compare-120461

Hengl, T. (2009). A practical guide to geostatistical mapping.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518. Retrieved from 10.7287/peerj.preprints.26693v2

Hiemstra, P., Pebesma, E., Twenhofel, C., & Heuvelink, G. (2008). Real-time automatic interpolation of ambient gamma dose rates from the dutch radioactivity monitoring network. *Computers Geosciences*. (DOI: http://dx.doi.org/10.1016/j.cageo.2008.10.011)

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight*, *4*. Retrieved from https://robjhyndman.com/papers/foresight.pdf

Izadi, G., Zhong, M., LaFollette, R. F., et al. (2013). Application of multivariate analysis and geographic information systems pattern-recognition analysis to production results in the Bakken light tight oil play. In *SPE hydraulic fracturing technology conference.* doi: 10.2118/163852-MS

Jaiswal, S. (2018). *K-means clustering in R tutorial.* Retrieved from https://www.datacamp.com/community/tutorials/k-means-clustering-r

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R.* Springer. Retrieved from https://faculty.marshall.usc.edu/gareth-james/ISL/

JJ. (2016). *MAE and RMSE — which metric is better?* Retrieved from https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

Kalogirou, S., & Georganos, S. (2019a). *Package "SpatialML".* CRAN. Retrieved from https://cran.r-project.org/web/packages/SpatialML/SpatialML.pdf

Kalogirou, S., & Georganos, S. (2019b). SpatialML: Spatial machine learning [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=SpatialML (R package version 0.1.3)

Koehrsen, W. (2018). *Hyperparameter tuning the random forest in Python.* Retrieved from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Kuhn, M. (2019). *The caret package.* Retrieved from https://topepo.github.io/caret/index.html

Kuhn, M. (2020). caret: Classification and regression training [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=caret (R package version 6.0-86)

Leroux, C. (2019). *Variogram and spatial autocorrelation.* aspexit. Retrieved from https://www.aspexit.com/en/variogram-and-spatial-autocorrelation/

#Fitting_a_variogram_model_to_the_data

Li, Y., & Han, Y. (2017, Nov 07). Decline curve analysis for production forecasting based on machine learning. In (p. 14). Kuala Lumpur, Malaysia: Society of Petroleum Engineers. Retrieved from https://doi.org/10.2118/189205-MS doi: 10.2118/189205-MS

Liaw, A., & Wiener, M. (2002a). Classification and regression by randomforest. *R News*, *2*(3), 18-22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Liaw, A., & Wiener, M. (2002b). Classification and regression by randomForest. *R News*, *2*(3), 18-22. Retrieved from https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Lichtenstern, A. (2013). Kriging methods in spatial statistics.

Lolon, E., Hamidieh, K., Weijers, L., Mayerhofer, M., Melcher, H., & Oduba, O. (2016). Evaluating the relationship between well parameters and production using multivariate statistical models: A Middle Bakken and Three Forks case history. In *SPE hydraulic fracturing technology conference.* doi: 10.2118/179171-MS

Longman, M., Luneau, B., & Landon, S. (1998). Nature and distribution of Niobrara lithlologies in the cretaceous western interior seaway of the Rocky Mountains region. *The Mountain Geologist*, *35*(04), 137-170. Retrieved from http://archives.datapages.com/data/rmag/mg/1998/longman.htm

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R.* CRC Press. Retrieved from https://geocompr.robinlovelace.net/spatial-cv.html

Luo, G., Tian, Y., Bychina, M., & Ehlig-Economides, C. (2018). Production optimization using machine learning in Bakken shale. In *Unconventional resources technology conference, Houston, Texas, 23-25 July 2018* (pp. 2174–2197). doi: 10.15530/URTEC-2018-2902505

Manfreda, J. (2015). *The origin of fracking actually dates back to the civil war.* Retrieved from https://www.businessinsider.com/the-history-of-fracking-2015-4?r=US&IR=T

Markets Insider. (2020). Retrieved from https://markets.businessinsider.com/commodities/oil-price?type=wti

Marko, K., Al-Amri, N. S., & Elfeki, A. M. (2014). Geostatistical analysis using GIS for mapping groundwater quality: case study in the recharge area of Wadi Usfan, western Saudi Arabia. *Arabian Journal of Geosciences*, *7*(12), 5239–5252. doi: 10.1007/s12517-013-1156-2

McDonnell, T. (2020). *The collapse of the US fracking industry, in seven charts.* Retrieved from https://qz.com/1830456/how-the-coronavirus-is-disrupting-the-us-fracking-industry/

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815. doi: 10.1016/j.ecolmodel.2019.108815

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, *94*(2), 284–289. Retrieved from http://www.jstor.org/stable/3693985

Mohaghegh, S., Gaskari, R., & Maysami, M. (2017). Shale analytics: Making production and operational decisions based on facts: A case study in Marcellus shale. In *SPE hydraulic fracturing technology conference and exhibition.* doi: 10.2118/184822-MS

Molnar, C. (2019). *Interpretable machine learning.* Retrieved from https://christophm

.github.io/interpretable-ml-book/

Montgomery, J., & O'Sullivan, F. (2017). Spatial variability of tight oil well productivity and the impact of technology. *Applied Energy*, *195*, 344–355. doi: 10.1016/j.apenergy.2017.03.038

Parr, T., Turgutlu, K., Csiszar, C., & Howard, J. (2018). *Beware default random forest importances.* explained.ai. Retrieved from https://explained.ai/rf-importance/index.html#4

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers Geosciences*, *30*, 683-691.

PSAC. (n.d.). *Fracking explained.* Petroleum Services Association of Canada. Retrieved from https://oilandgasinfo.ca/all-about-fracking/fracking-explained/

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rapier, R. (2017). How the shale boom turned the world upside down. *Forbes*.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*. Retrieved from https://arxiv.org/abs/1811.12808

Reuters. (2019). *U.S. shale oil output to rise to record 8.52 million barrels per day in July: EIA.* Retrieved from https://www.reuters.com/article/us-usa-oil-productivity/u-s-shale-oil-output-to-rise-to-record-8-52-million-barrels-per-day-in-july-eia-idUSKCN1TI2DP

Rigzone. (n.d.). *How does casing work?* Retrieved from https://www.rigzone.com/training/insight.asp?insight_id=333&c_id=

Rystad Energy. (2017a). Halliburton outperforms the competition in Bakken: mixed empirical evidence. *Rystad Energy Report: Unpublished*.

Rystad Energy. (2017b). Permian Delaware solution: Wolfcamp, Schlumberger and Cimarex Energy. *Rystad Energy Report: Unpublished*.

Rystad Energy. (2017c). Permian Midland: complexity of optimal proppant intensity, degradation in infill well productivity and PXD's systematic outperformance. *Rystad Energy Report: Unpublished*.

Rystad Energy. (2019). Updated regression model confirms child well underperformance. *Rystad Energy Report: Unpublished*.

Satter, A., & Iqbal, G. (2015). *Reservoir engineering: The fundamentals, simulation, and management of conventional and unconventional recoveries.*

Schlumberger. (n.d.). *Permeability.* Schlumberger Oilfield Glossary. Retrieved from https://www.glossary.oilfield.slb.com/en/Terms/p/permeability.aspx

Schuetter, J., Mishra, S., Zhong, M., & LaFollette, R. (2015, Jul 20). Data analytics for production optimization in unconventional reservoirs. In (p. 20). San Antonio, Texas, USA: Unconventional Resources Technology Conference. Retrieved from https://doi.org/10.15530/URTEC-2015-2167005 doi: 10.15530/URTEC-2015-2167005

Scornet, Erwan. (2017). Tuning parameters in random forests. *ESAIM: Procs*, *60*, 144-162. doi: 10.1051/proc/201760144

Seg Wiki. (2020a). Retrieved from https://wiki.seg.org/wiki/Acquisition#Downfalls_of_Land_Acquisition

Seg Wiki. (2020b). Retrieved from https://wiki.seg.org/wiki/Play_fair_analysis

Shah, T. (2017). *About train, validation and test sets in machine learning.* Retrieved from https://towardsdatascience.com/train-validation-and-test

-sets-72cb40cba9e7

Shortridge, A. (2019). *Spatial data analysis: Ordinary kriging.* Retrieved from https://msu.edu/~ashton/classes/866/papers/gatrell_ordkrige.pdf

SM Energy Company. (2015). *Hydraulic fracturing process.* Retrieved from https://www.youtube.com/watch?v=T_yfPcX1gG4&t=112s

Speight, J. G. (2017). Chapter three - gas and oil resources in tight formations. In J. G. Speight (Ed.), *Deep shale oil and gas* (p. 121 - 174). Boston: Gulf Professional Publishing. doi: 10.1016/B978-0-12-803097-4.00003-6

Statista. (2020). Retrieved from https://www.statista.com/statistics/271823/daily-global-crude-oil-demand-since-2006/

Statoil. (2013). *Drilling hydraulic fracturing - how it's done, responsibly.* Retrieved from https://www.equinor.com/content/dam/statoil/documents/united-states/Shale_DrillingHydraulicFacturing.pdf

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008, 11). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. doi: 10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007, 25). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. doi: 10.1186/1471-2105-8-25

Swalin, A. (2018). *How to handle missing data.* Retrieved from https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Thul, D. (2012). *Niobrara source rock maturity in the Denver Basin: A study of differential heating and tectonics on petroleum prospectivity using programmed pyrolysis.* Colorado School of Mines. Retrieved from https://books.google.no/books?id=K8lImwEACAAJ

Trading Economics. (2020). Retrieved from https://tradingeconomics.com/united-states/crude-oil-rigs

Veness, C. (n.d.). *Calculate distance, bearing and more between latitude/longitude points.* Retrieved from http://www.movable-type.co.uk/scripts/latlong.html

Wackernagel, H. (2013). *Multivariate geostatistics: an introduction with applications.* Springer Science & Business Media.

Wagener, D. V. (2018). *Oil and natural gas resources and technology.* U.S. Energy Information Administration. Retrieved from https://www.eia.gov/outlooks/aeo/grt.php

Wang, S., & Chen, S. (2019). Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. *Journal of Petroleum Science and Engineering*, *174*, 682–695. doi: 10.1016/j.petrol.2018.11.076

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists.* John Wiley & Sons.

Xie, J., Lee, S., Wen, X.-H., & Wang, Z. (2013). Uncertainty assessment of production performance for shale gas reservoirs.. doi: 10.2523/16866-ABSTRACT

Zhong, M., Schuetter, J., Mishra, S., Lafollette, R. F., et al. (2015). Do data mining methods matter?: A Wolfcamp Shale case study. In *SPE hydraulic fracturing technology conference.* doi: 10.2118/173334-MS

Zhou, Q., Kleit, A., Wang, J., & Dilmore, R. (2014). Evaluating gas production performances in Marcellus using data mining technologies. In *Unconventional resources technology conference, Denver, Colorado, 25-27 august 2014* (pp. 20–36). doi: 10.1016/j.jngse.2014.06.014

Appendix

A1 Appendix A

This appendix presents output from the discarded approach that involved the construction of predictive models where both geological and well-design variables were included as predictors.

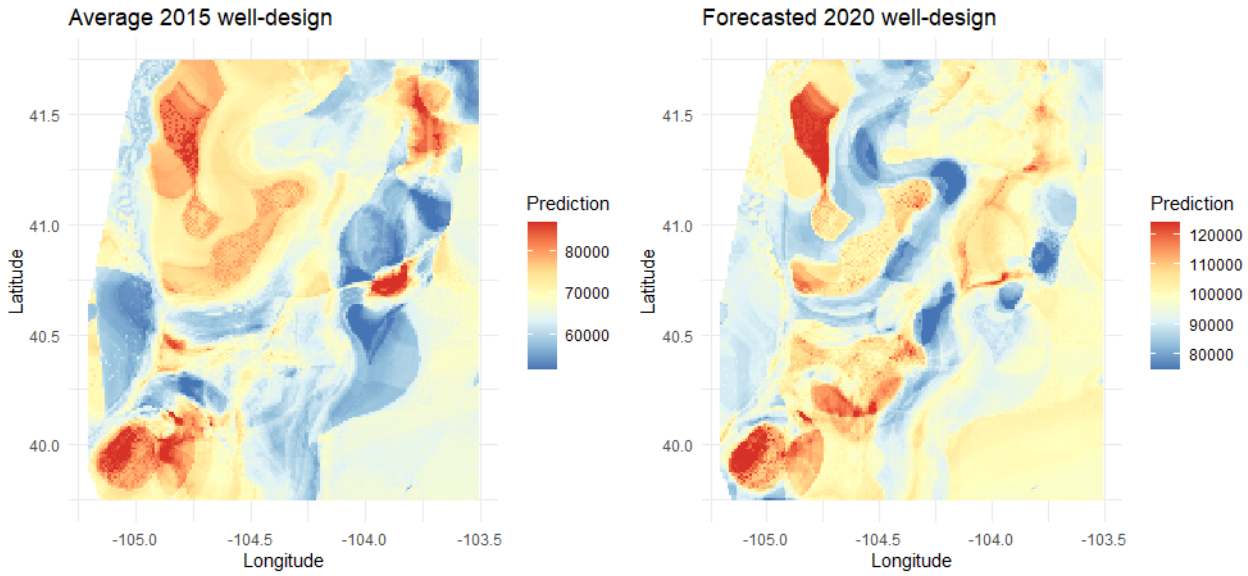

Figure A1.1. Predicted production volumes when well-design variables are included along geological variables. Left: Well-design is set to average 2015 levels. Right: Well-design is set to forecasted 2020 levels.

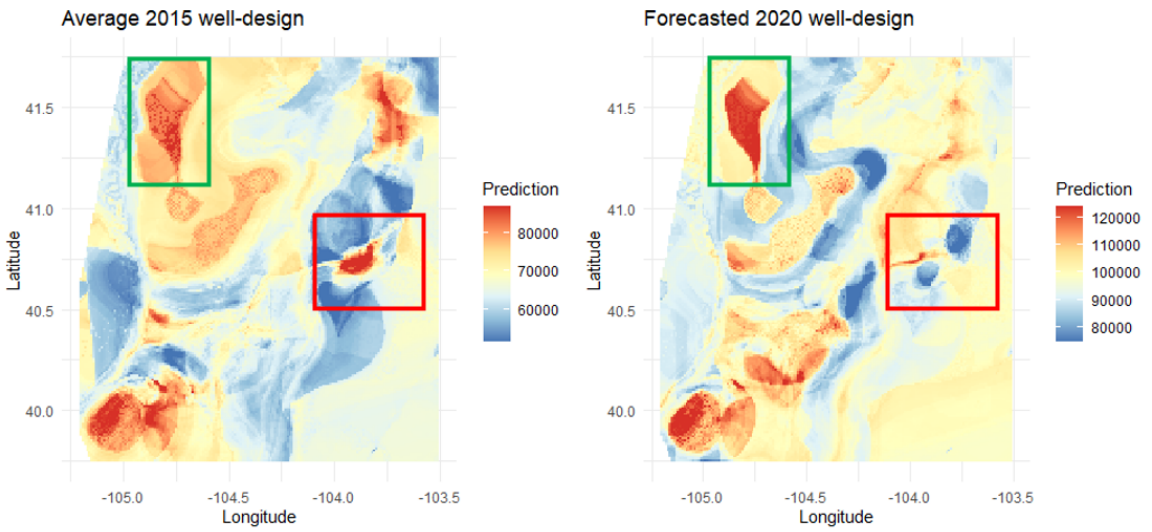

Figure A1.2. Modified version of Figure A1.1. Green square: Area where predicted production volume improves with increased well-design. Red square: Area where predicted production volume is essentially unaffected by increased well-design.