SWEET-SPOTTING FOR UNCONVENTIONALS PHASE I PROJECT REPORT

FRANZ J KIRÁLY
DEPARTMENT OF STATISTICAL SCIENCE, UNIVERSITY COLLEGE LONDON

1 SUMMARY

Sections 1.1 - 1.5 provide a quick summary on the results of phase I. Sections 2 provides an extended summary and discussion. The subsequent sections give a technical overview on the main findings.

1.1 MAIN OBJECTIVES

The primary goal of the project is to determine whether, and if yes, how and how well, the location of "sweet spots" in unconventional reservoirs can be predicted by using geological and operational features from the data sets provided.

The main goal of **Phase I** (Jan – Mar 2015) was to obtain a basic understanding on the data sets and their relation to each other. It is planned that **Phase II** (Mar – Jul 2015) will focus on (mainly spatial) integration of the data sets, **phase III** (Jul 2015- Jan 2016) on evaluation and validation of predictive models.

1.2 SHORT OVERVIEW OF DATA SETS PROVIDED

Three data sets, describing features relating to the Eagle Ford Formation, were provided for initial analyses:

(1) a data set containing mainly production logs and operational features of producing oil and gas wells (N= 11447), obtained from *IHS Energy*, in the following called the **IHS data set**. The primary goal of the project is predicting production of wells similar to those in the IHS data.
(2) a data set containing mainly geological features from core samples, obtained from *Core Laboratories*, in the following called the **Core Labs data set**. The samples are available at only a few locations, the so-called core wells (N=88).
(3) a data set obtained from integrating parts of the IHS and the Core Labs data sets, used in prediction models previously obtained in a project with *Kaggle*. Most notably, it contains integrated features, and predictions obtained from two models; in the following, it will be called the **Kaggle data set**.

All three data sets consist of multiple, related data tables. The main feature of the IHS and especially the CoreLabs data sets are a large number of variables (several hundreds) associated with different geographic (core and well) locations in the Eagle Ford Shale. The data sets were used and will be discussed as provided by the *Shell Computation and Modeling Group (SIEP-PTI/RP)* at the beginning of the project. More details on the data sets and their relation can be found in section 3.

1.3 MAIN QUESTIONS

The main goal of phase I was familiarization with the data sets and general exploration. Furthermore, different scientific questions were investigated, which can be roughly summarized as follows:

- ***(A) Which target/outcome variable(s) for well production is/are most appropriate for the prediction? What are important features of these target/outcome variables?***
- ***(B) Which covariates are relevant to predict from? What are important features of these?***
- ***(C) More generally: what are important aspects of the data sets that need to be addressed when building a predictive model?***
- ***(D) Checking claims made in the internal state-of-the-art, mostly consisting of the rules-based prediction and the Kaggle models, regarding (A), (B), (C), and the goodness of the existing predictors.***

Scientific details regarding these questions and the state-of-the-art is described in sections 2.1 and 2.2.

1.4 SUMMARY OF RESULTS

Analyses were focused on part of the data sets due to time constraints and size of the data sets. Most are work in progress, and may change over the course of the project.

(A) Using first 12 months of production is not unreasonable. Kaggle's normalization of this quantity by perforation depth is questionable and possibly problematic. There is evidence that longer-term production may be captured by using adaptive spectral summaries instead. A final, quantitative answer can only be given when a complete prediction model is in place.

(B) Kriging production (= using just longitude/latitude) is already very good if production for a large number of wells (>1000) is available. There is evidence, obtained from studying Kaggle's interpolates, that adding in geological variables helps when some but not too few production wells (ca 100-1000) are available. Some of the covariates used by big rules and Kaggle appear useful. A final answer can only be given in the context of a complete prediction workflow. Ideas from parameter tuning may be useful in selecting variables automatically.

(C) Integrating features between core and production well locations is crucial. Multiple cores (ca 10-100) can be present at a single core well, so the integration of core wells vs core samples is crucial as well. Interpolation and aggregation are methods that can address these separately. There are methods which can combine either or both of the two integration steps with prediction. These and related questions will be investigated in more detail in phase II of the project.

(D) Some central claims made by Kaggle could not be validated against scientific standards and therefore should be treated with caution. Since the Kaggle models cannot be trained and provide only predictions, there is no scientific way of checking the goodness of the models themselves. The predictions made by the Kaggle "rules based" model are as good as random guesses in most metrics of goodness (including Kaggle's). The predictions made by the final Kaggle model are better than random, but as good as interpolating production without looking at the geological variables; also, the Kaggle predictions are outperformed by un-tuned off-the-shelf methods in standard metrics of goodness (including Kaggle's) and validation setups (including Kaggle's) for which experiments were set up.

References to the analyses from which the above statements can be derived are listed in Section 2.3.

1.5 MAIN RECOMMENDATIONS

(1) In the continuation of this project, top priority should be assigned to investigating interpolation and integration aspects of the data sets, as initially planned for Phase II.
(2) If resources are available, a complete exploration and cleaning of the data sets may be beneficial before reaching phase III.
(3) No decisions should be based solely on claims or software by Kaggle while the scientific and methodological problems in their work persist, and until proper, reproducible evidence in their favour is provided.

1.6 RESOURCE ALLOCATION

Time constraints and the size of the data sets did not allow for a detailed study of all data provided, therefore parts known or suspected to be relevant were selected and investigated with the question "which variables are important" (see question (A) and (B)) in mind.

A considerable amount of time was also spent on understanding and cleaning the data, since despite commendable efforts on the sides of the Computation and Modeling group, documentation of the IHS and Core Labs data sets remained to problematic extent obscure, and data errors were prevalent in many parts; both mostly attributable to the state in which data and documentation was provided by IHS and Core Labs. The Kaggle data set was problematic as well, due to almost complete lack of information regarding methodology in the available reports and the fact that no code was provided whatsoever.

Some time was spent on investigating questions related to integration, interpolation and variable selection, to understand the main difficulties which will be faced in Phases II and III (see question (C)).

1.7 REPOSITORY

Reproducible code of all analyses, supplementary material, and project documentation (including this report) are available in an SVN repository managed by the Computation and Modeling Group at Shell Research and accessible via the URL: https://shellcompmod.svn.cloudforge.com/exchangekiraly/.

References to files and directories below are relative to the root/"trunk" folder of that repository, in its state of March 16, 2015.

2 ABOUT THE MAIN QUESTIONS

2.1 DISCUSSION AND SCIENTIFIC CONTEXT

(A) ***Which target/outcome variable(s) for well production is/are most appropriate for the prediction?***

This is the most important question, since to predict one must know *what* to predict.

There is no purely quantitative way to answer this question from the given data sets, because whether it makes sense to predict a particular variable depends on what such a prediction would be used for; which most likely will involve economic or technical considerations beyond the given data sets.

However, for a given target variable, and a given model, there are standard quantitative ways to check how well the given model is in predicting the given target variable in a given setting. For such an assessment, the target variable needs to figure in the predictive model.

What are important features of these target/outcome variables?

This is a related, open question, which can be studied by exploring the target variable and investigating it in predictive models in which it is part of.

(B) Which covariates are relevant to predict from?

This is the second important question, addressing *from what* to predict.

Naturally, it requires an answer to question (A), i.e., the choice of prediction target, and it will in general heavily depend on the choice of a target variable.

Once a target variable has been chosen, there are standard quantitative ways to check how well a given model, which includes the choice of covariates, predicts the given target variable in a given setting. This allows to investigate which choices of covariates are good for prediction in a given type of model, though the particular way to do so is, for most existing types of models, a subject of research called “variable selection”.

For these types of assessment, both target variable and covariates need to figure in a predictive model.

What are important features of these?

This is a related, open question, which can be studied by exploring the covariates, their relation to the target variable and each other, and investigating predictive models which both covariates and the target are part of.

(C) More generally: what are important aspects of the data sets that need to be addressed when building a predictive model?

Since there are multiple data sets containing multiple tables, this question is crucial in understanding how information from all sources can be integrated.

A basic yet very important sub-question is the database-type relation between the data sets, i.e., which unique identifiers occur where and how they are linked across tables; this is straightforward to answer given sufficient documentation of the data sets.

The question also contains a more open aspect regarding the statistical relationship of the different data sets, which can be studied in an explorative manner and, more quantitatively, in relation to given predictive models that are built across data sets.

(D) Checking claims made in the internal state-of-the-art, mostly consisting of the rules-based prediction and the Kaggle models, regarding (A), (B), (C), and the goodness of the existing predictors.

This is a crucial question since the state of the art is both an important validation baseline and potentially contains partial answers to questions (A), (B), (C).

Claims that are scientifically precise and methodologically sound can be quantitatively evaluated in simulation experiments involving the data sets.

Once a prediction model is available as a black box that allows (a) "training"/"learning" the model on a training set and (b) making predictions on unseen data, it can be evaluated with respect to various standard measures of goodness relating to the error of prediction, yielding a concise description of how well the model performs as a sum of its parts, which includes selection of covariates, target outcome, and choice of model type.

2.2 STATE OF THE ART

The following describes the state of the art at project start.

(A) ***Which target/outcome variable(s) for well production is/are most appropriate for the prediction?***

(A.i) There is currently no internal standard and no (successful) outcome used in scientific literature on unconventionals so far.

(A.ii) A very desirable outcome to predict would be EUR (=estimated ultimate recovery)/oil production and EUR/gas production, or, alternatively, the location where EUR would be high (as opined by the members of the Computation and Modeling group in the 1st PDA meeting). However, EUR is not available in the raw data for many wells due to censoring, i.e., since many wells are still actively producing oil/gas.

(A.iii) In their models, Kaggle predicted the categorical variable of being in the top quartile of first-12-months-production oil/gas, divided ("normalized") by length of well perforation. No proper justification is given in their reports for choosing this outcome over alternatives, in particular for (a) why the quartile is chosen as compared to other percentiles, (b) why the binary outcome of being above a percentile is chosen as opposed to predicting production itself, (c) the division by length of well perforation. Especially (c) is a questionable choice, since perforation length becomes only known *after* drilling the well. While each of (a)-(c), as a separate choice may be justified, the effect of combining them may be unpredictable and is thus problematic without further evidence for this choice.

What are important features of these outcomes?

Existing claims did not directly apply to this project because of different project objectives (unconventionals), or had to be considered as unvalidated for the time being due to lack of reproducible code.

(B) ***Which covariates are relevant to predict from?***

(B.i) In scientific literature, sweet-spotting in unconventionals remains an open problem.

(B.ii) There is an internal scoring scheme called the "big rules", which is given the values of certain geological variables and outputs a score (not a prediction of production). According to discussion with Jan Limbeck, this score is then in practice used to reject poor locations (screen out tool) rather than to find the actual sweet spots of a prospect since predictions of upsides have turned out to be unreliable in the past.
There is also a model built by Kaggle, inspired by the "big rules", which uses a similar set of covariates and outputs a score (not a prediction of production).

There was no reproducible quantitative evidence on why those covariates are useful.

(B.iii) Kaggle have presented a machine learning model for predicting production which uses a larger number of covariates as inputs. The principal quantitative claim made by Kaggle is that their black box model outperformed their "big rules" model in terms of predicting whether a production well is in the top quartile of first-12-months production divided by perforation well length.

This claim was checked and verified prior to project start by Mingqi Wu for the output made by the black box predictor provided by Kaggle. However, since the black box model cannot be trained, and since Kaggle had access to large parts of the data, it cannot be checked whether those outputs are really predictions.

Therefore, there was no reproducible quantitative evidence on why the covariates selected by Kaggle were useful.

What are important features of these?

Existing claims did not directly apply to this project because of different project objectives (unconventionals), or had to be considered as unvalidated for the time being due to lack of reproducible code.

(C) ***More generally: what are important aspects of the data sets that need to be addressed when building a predictive model?***

The exact data-base type relation between the IHS and Core Labs data sets is known and given by many-to-many connections given by unique well identifiers and core sample identifiers.

The wells in the Kaggle data set can be matched to a subset of wells in the IHS data set. There is no information on how Kaggle selected the wells in their data set. Though there are very high-level descriptions in the Kaggle report, there is no reliable information or how the integrated/interpolated covariates in the Kaggle data set were obtained.

Regarding statistical features, various claims are documented in existing project materials with Shell Research.

All claims remain unvalidated and questionable for the purpose of this project, since reproducible code outside the scope of the present project was not available.

(D) ***Checking claims made in the internal state-of-the-art, mostly consisting of the rules-based prediction and the Kaggle models, regarding (A), (B), (C), and the goodness of the existing predictors.***

Prior to project start, Mingqi Wu has checked the principal quantitative claim made by Kaggle, which was that their black box model outperformed their "big rules" model in terms of predicting their top quartile outcome.

However, since the black box model cannot be trained, and since Kaggle had access to large parts of the data, it cannot be checked whether the output of the Kaggle black box model are really predictions. Thus Kaggle's claim that their model is useful in predicting their outcome remains unvalidated and possibly uncheckable.

Therefore, there was no reproducible quantitative evidence regarding questions (A), (B), (C) in the work of Kaggle. There was also no further such evidence from the other projects due to lack of code.

2.3 EXTENDED SUMMARY

(A) ***Which target/outcome variable(s) for well production is/are most appropriate for the prediction? What are important features of these target/outcome variables?***

Partial answers are provided in sections 4.1-4.3. The spectral summaries presented in 4.3 can be learnt from training data and may therefore be advantageous to consider as opposed to non-adaptive summaries.

The investigations in sections 4.2 show that depth variables may be informative, but exploration alone provides no evidence on whether incorporating depth-related variables, as proposed by Kaggle, is useful.

As explained there, quantitative answers on usefulness of outcome variables can only be obtained in the context of a complete predictive model.

(B) ***Which covariates are relevant to predict from? What are important features of these?***

Sections 5.2-5.4 contain partial answers in the experiments presented there.

The results in tables 5.2.1 and 5.3.1 show that predictions that are significantly better than random guesses can be made from the interpolated variables in both Kaggle data sets.

Section 5.4 contains evidence on situations in which adding geological variables above well position may be useful.

As explained there, quantitative answers on usefulness of outcome variables can only be obtained in the context of a complete predictive model – which here, in particular, means in a model where the interpolated variables have been obtained in reproducible ways.

(C) ***More generally: what are important aspects of the data sets that need to be addressed when building a predictive model?***

The structural/database type aspects of this are outlined in section 3.1. Statistical aspects of this question will be studied in phase II.

(D) ***Checking claims made in the internal state-of-the-art, mostly consisting of the rules-based prediction and the Kaggle models, regarding (A), (B), (C), and the goodness of the existing predictors.***

The fact that Kaggle's rule based model is as good as a random guess in predicting top percentiles of normalized oil production is contained is implied by figure 5.2.2 (the yellow curve is very close to the red "random guess" baseline). A similar, more quantitative result using the rank versions RMSE and MAE is available in the referenced source code.

The fact that Kaggle's black box model is as good as interpolating production from 131 production wells, without looking at the geological variables obtained from the Core Labs data, is part of table 5.3.1. Results for various simple off-the-shelf models outperforming Kaggle's black box model in different validation setups are listed in sections 5.2 and 5.3.

The recommendation to treat Kaggle's claims with caution is based on the observation that their claims partly contradict the above findings, and that they did not present nor reproducibly describe a validation setup for their claims in their phase I and phase II report, without which many of their claims remain uncheckable.

3 DESCRIPTION OF THE DATA SETS

3.1 RELATION OF THE DATA SETS

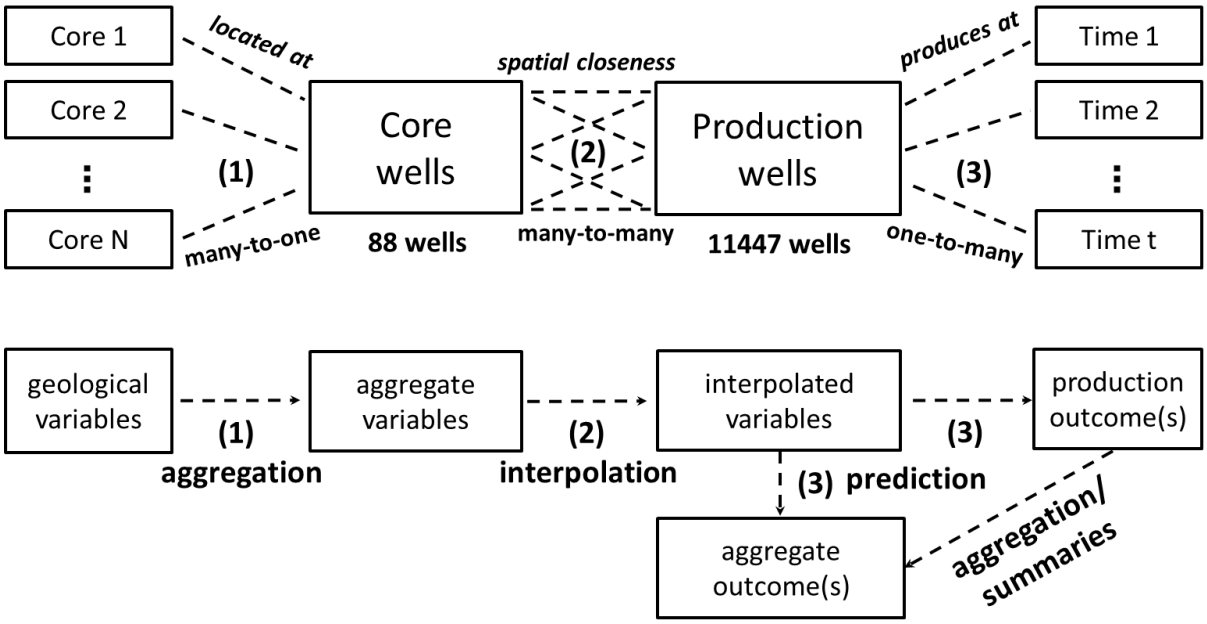

Figure 3.1.1: schematic summary of the relation between the production wells in the IHS data set, and core wells and core samples in the Core Labs data sets. Geological variables are available per core sample; (1) multiple core samples can be located at a single core well. (2) the production wells are distinct from the core wells but are located in the same region, the Eagle Ford shale. (3) each production well produces a certain amount of gas/oil in a certain time interval. The main task is to investigate whether predictions of production outcomes for the production wells can be made and/or improved by using the geological information in the core samples.

Figure 3.1.1 provides a schematic overview on how the Core Labs and the IHS data set relate to each other and the main task of investigating prediction methods for well production outcomes. Figure 3.1.2 shows the core wells in the Core Labs data set and the production wells in the IHS data set on a map. The principal features and difficulties in relating the two data sets to each other are:

(1) Multiple core samples, in the order of 10-100, are available per core well, in varying numbers.
(2) the core wells only partly overlap with the production wells. Measurements common for the production wells are not available for the core wells and vice versa.
(3) Production of oil/gas is available as production per time intervals of which there are in the order of 10-100 per production well.

It is an open question what the best way to address these features in the IHS and Core Labs data sets is. A straightforward approach would be to deal with (1), (2) and (3) separately – for example by (1) variable aggregation to obtain aggregate variables for the core wells from the core samples, and (2) interpolation to obtain interpolated variables at the production wells from those at the core wells, then predicting (3) summaries or aggregates for the production outcomes from the interpolated variables. More sophisticated approaches could deal with several of (1), (2), (3) at once, such as latent variable models for (2) & (3), or distributional interpolation for (1) & (2), though such approaches need not predict better.

More generally, it remains to be investigated – and quantified in a predictive setting – which methods of integration are best.

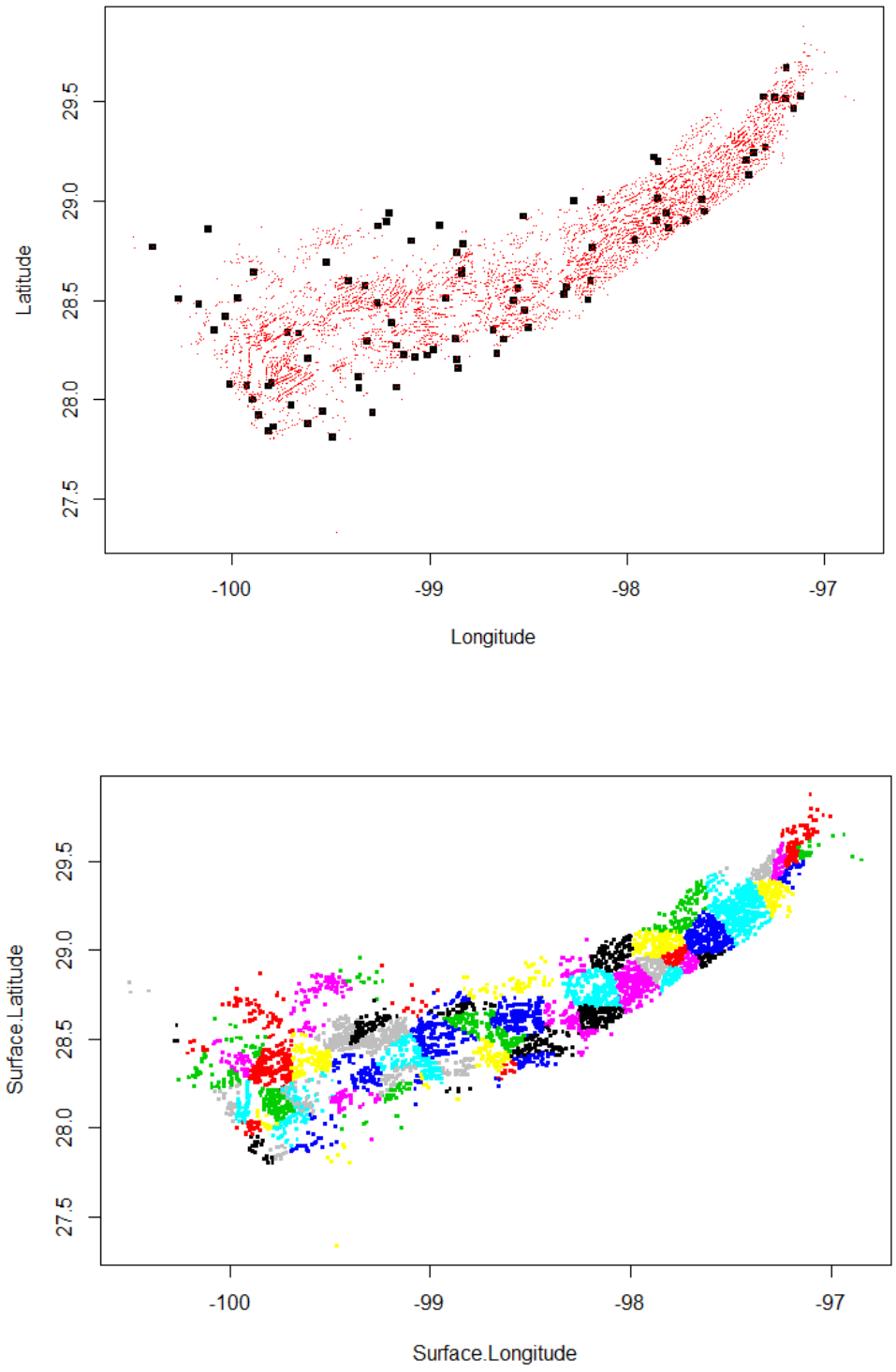

Figure 3.1.2: maps of the wells in the IHS and Core Labs data set (after cleaning and outlier removal). The top map shows the positions of the production wells in the IHS data set (small red dots) and the core wells in the Core Labs data set (big black dots). The bottom map shows the IHS wells, coloured in clusters given by the core well closest in the Euclidean distance of longitude/latitude (colours may repeat).

3.2 THE IHS DATA SET

The IHS data set contains two sets of tables obtained from the tabs of two excel sheets: the production workbook and the well workbook. Tables 3.2.1 and 3.2.2 contain information on both, given the current stage of the project.

The basis for the investigation of the production variables in section 4 are the production header, production well, and monthly production tables.

name of table	contains	rows	unique IDs	Remarks
Production header	general well information	11447	11447 Ent.	
Production well	location and depths	11447	11447 Ent.	
Production test	well test data	24939	9773 Ent.	not looked at (time constraints)
Monthly production	monthly production data	211704	10858 Ent.	production raw data
Production abstract	production summaries	10858	10858 Ent.	obtainable by aggregating the
Annual production	annual production data	27040	10858 Ent.	monthly production; IP-cum-norm
IP-cum-norm values	production summaries	10944	10944 Ent.	contains slightly more Entities

Table 3.2.1: overview of the IHS production workbook. Each of the data tables in the production workbook is summarized by: the type of information it contains, the number of rows before cleaning, the number of unique production well IDs ("Entities") in the table, and remarks regarding the table

name of table	contains	rows	unique IDs	Remarks
Well header	general well information	10683	10683 API	
other tables				not looked at (time constraints)

Table 3.2.2: overview of the IHS well workbook. Each of the data tables in the well workbook is summarized by: the type of information it contains, the number of rows before cleaning, the number of unique production well IDs ("APIs") in the table, and remarks regarding the table

Data cleaning, outlier removal, and exploration is documented in-code.

Repository reference:

dataAnalysis/IHS - merged logs exploration/IHSframe_prune_rare.R
dataAnalysis/IHS - merged logs exploration/exploration_IHS_merged.R
dataAnalysis/IHS - merged logs exploration/IHSdata_master.R (master file)

3.3 THE CORE LABS DATA SET

The Core Labs data set contains multiple tables: the production workbook and the well workbook. Table 3.3.1 contains information on both, given the current stage of the project. Entities in the IHS data set can be directly linked to UWIs in the Core Labs data set. The table currently under investigation is the Shale Gas Analysis table which should contain summaries or aggregates obtained from most other tables in the Core Labs data set.

	Contains	Rows	unique IDs	Remarks
Data Inventory	list of core samples	1773	89 UPWIs	contains list of all core samples
Shale Gas Analysis **(Aggregated Data)** **(+ Processed Log Data)**	huge list of geological covariates, including the ones used by Kaggle	4406 3972 2357	88 UWIs 89 UPWIS (53/54 in A&PL)	there are 3 versions of this data set with different variables. Most complete is S.G.A.
Routine Core Analysis **Routine Core Analysis API**	few geological variables	4411 32984	89 UPWIs 486 APIs	mostly in Shale Gas Analysis mostly outside Texas/Eagle Ford
other tables				not looked at (time constraints) should be mostly in S.G.A.

Table 3.3.1: overview of the Core Labs tables. Each of the data tables in the production workbook is summarized by: the type of information it contains, the number of rows before cleaning, the number of unique well IDs in the table, and remarks regarding the table. Well IDs come in three forms, depending on the data set: UWIs ("Unique well identifier"), UPWIs ("Unique private well identifier") and APIs. The UWIs and UPWIs are one-to-one, with the exception of two distinct UPWIs corresponding to a single UWI.

Data cleaning, outlier removal, and exploration is documented in-code.

Repository reference:
dataAnalysis/CoreLabs - logs exploration/CoreLabs_cleaning_RoutineCoreAnalysis.R
dataAnalysis/CoreLabs - logs exploration/CoreLabs_cleaning_ShaleGasAnalysis.R
dataAnalysis/CoreLabs - logs exploration/CoreLabs_exploration_RoutineCoreAnalysis.R
dataAnalysis/CoreLabs - logs exploration/CoreLabs_exploration_ShaleGasAnalysis.R
dataAnalysis/CoreLabs - logs exploration/ CoreLabs_compare_3versionsof_ShaleGasAnalysis.R
dataAnalysis/CoreLabs - logs exploration/master_CoreLabs_explore.R (master file)

3.4 THE KAGGLE DATA SET

Two of the data sets obtained from the project with Kaggle were looked at for the present analyses, as described in table 3.3.1. The data set contains geological covariates which are already interpolated to production wells.

	Contains	Rows	unique IDs	Remarks
Rules features	Interpolated covariates for Kaggle's big rules predictor, and predictions	2631	2631 UWIs	
Eagleford Oil Input	Interpolated covariates for Kaggle's black box predictor, and predictions	2734	2631 UWIs	a few UWIs are duplicated

Table 3.3.1: overview of the Kaggle tables. Each of the data tables is summarized by: the type of information it contains, the number of rows before cleaning, the number of unique production well IDs in the table, and remarks regarding the table.

Exploration is documented in-code.

Repository reference:
dataAnalysis/CoreLabs - logs exploration/RuleFeatures_exploration.R

4 ANALYSIS OF THE IHS DATA SET

The IHS data set was looked at first since it contains the variables related to the outcomes and main question (A) – which production outcome is preferable. As explained in section 2.2, the question is not finally answered, therefore the production outcomes in the IHS workbook were studied in detail. This section contains an overview on the most important findings in this respect.

4.1 EXPLORATION OF PRODUCTION VARIABLES IN PRODUCTION WORKBOOK

The original format of oil and gas production in the IHS data are monthly production curves, which are then aggregated to obtain summaries such as first-12-month production, annual production, or yearly production, see section 3.2. Ultimate recovery is not available directly from the data set, since many wells are still producing. As a simple proxy for estimated ultimate recovery, first-12-months of production was available for most wells and was therefore studied first.

Figures 4.1.1 and 4.1.2 show histograms of first-12-months of gas and oil production for oil and gas producing wells. There is a large variation in the amount of produce, and most wells produce both gas and oil. As it can be seen in figure 4.1.2, the distributions are less skewed on the logarithmic production scale.

The labelling of a well as an “oil well” or a “gas well” is intrinsic to the IHS data set and has a qualitative influence on the shape of the empirical distribution.

Figures 4.1.3 and 4.1.4 show production color-plotted in geographical coordinates. Gas production shows a qualitatively smooth spatial distribution in comparison to oil production which shows variation on a smaller scale.

Furthermore, the map plots for first 12 months of production and the first singular score of production curves (see 4.2) are qualitatively similar.

In all plots, there are visible qualitative differences between wells labelled as “gas wells” and wells labelled as “oil wells”.

Conclusions which can be drawn from these exploratory observations:

(1) when considering production wells, it may be generally advised to consider the “gas wells” and “oil wells” separately
(2) it may be generally advised to consider gas production and oil production variables separately; especially, the relevant spatial features of gas and oil production variables may be on different scales or, more generally, of different kind
(3) considering production variables – as outcomes or covariates, in exploration or prediction – on a log-scale may be beneficial

Whether these observations are indeed relevant in the prediction task can only be validated and quantified in the context of a complete predictive model.

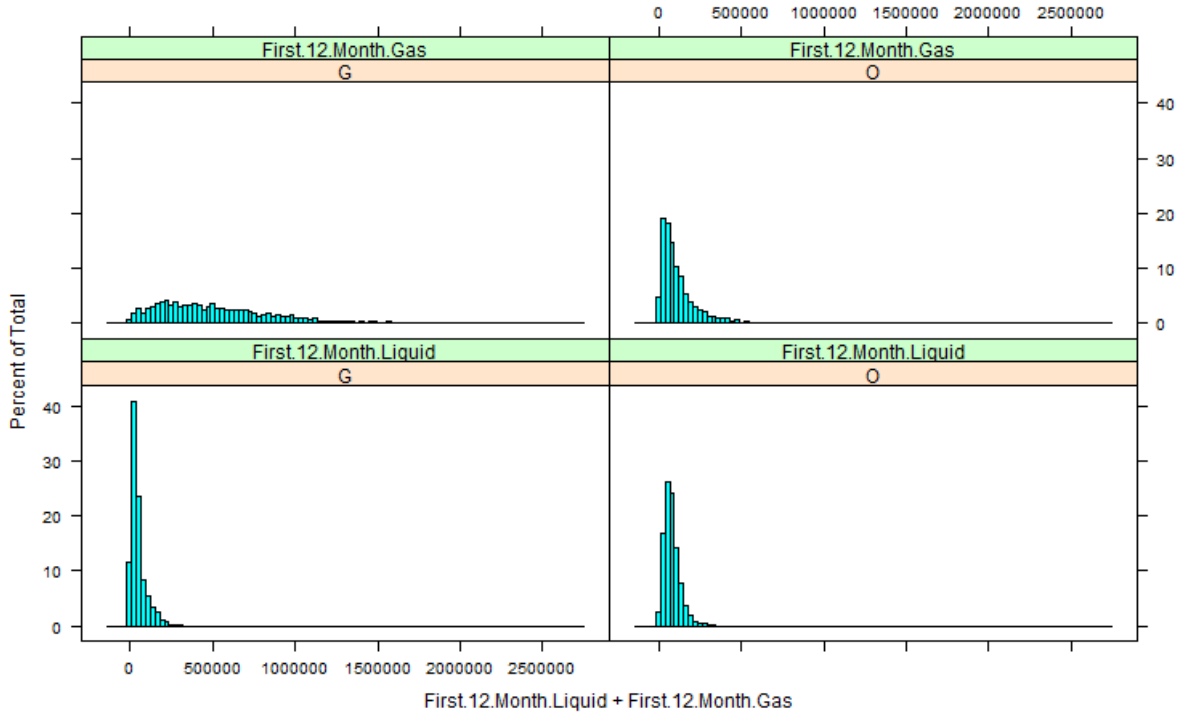

Figure 4.1.1: histograms of first twelve months gas (top two panels) and oil production/bbl (bottom two panels), stratified by whether the primary product of the well is gas (left two panels) or oil (right two panels)

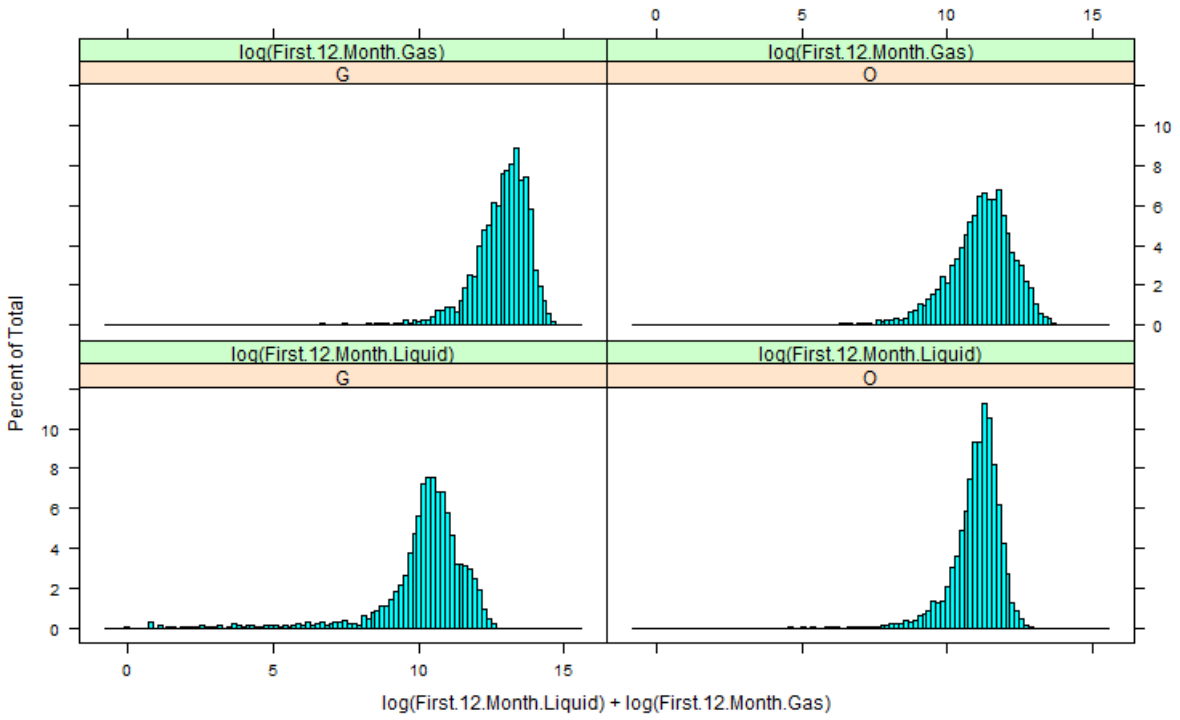

Figure 4.1.2: histograms of logarithm of first twelve months gas (top two panels) and oil production/bbl (bottom two panels), stratified by whether the primary product of the well is gas (left two panels) or oil (right two panels)

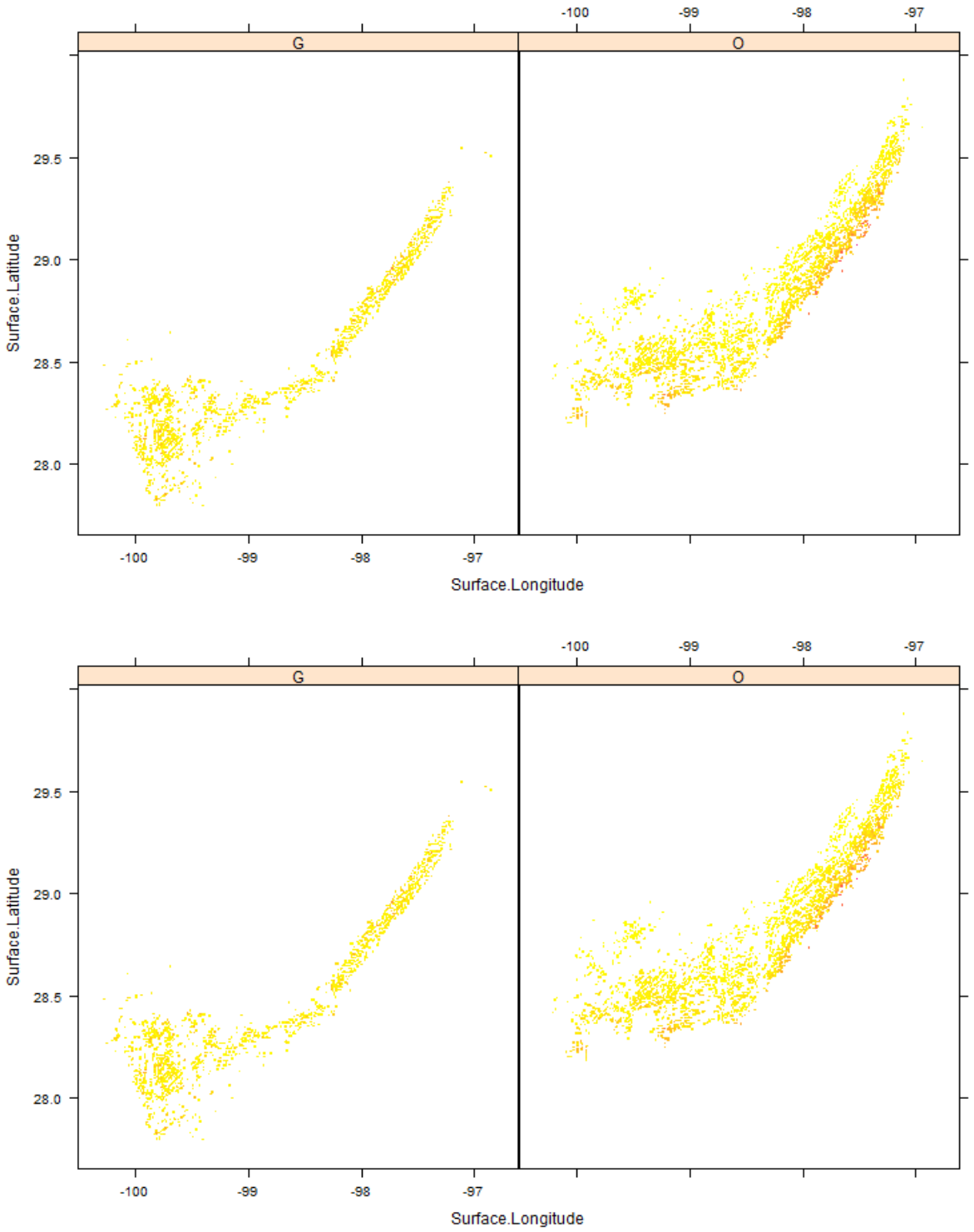

Figure 4.1.3: colored map plots of gas production, in units produced in first 12 months (top panel) and first singular score (bottom panel). Red = high production. Stratified by whether the primary product of the well is gas (left halves) or oil (right halves).

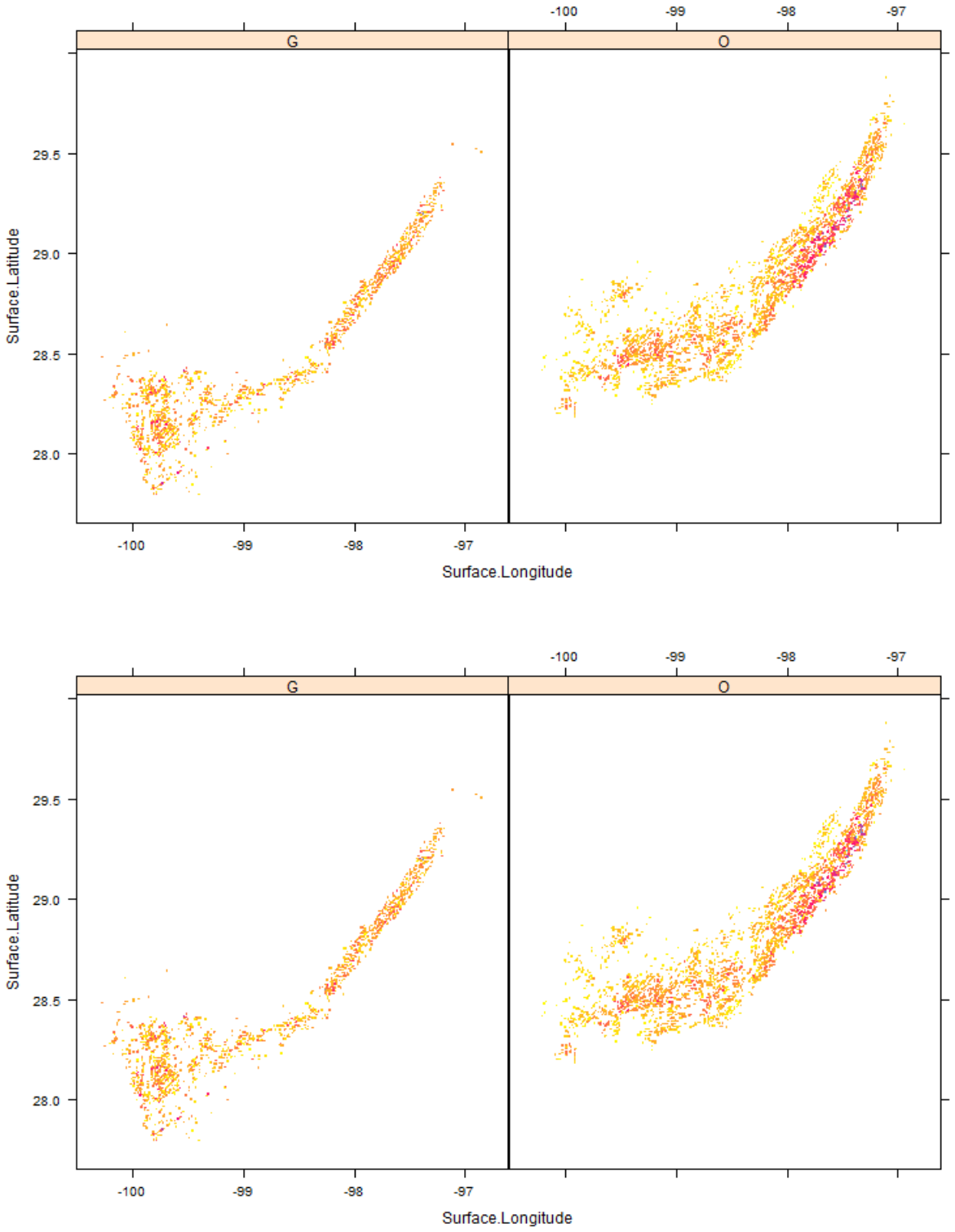

Figure 4.1.4: colored map plots of oil production, in units produced in first 12 months (top panel) and first singular score (bottom panel). Red = high production. Stratified by whether the primary product of the well is gas (left halves) or oil (right halves).

Repository reference:

dataAnalysis/IHS - merged logs exploration/exploration_IHS_merged.R

dataAnalysis/IHS - merged logs exploration/IHSdata_master.R (master file)

4.2 EXPLORATION OF DEPTH VARIABLES AND NORMALIZATION

One important question relating to (A) is whether considering the vertical structure of the drill hole and the geological layers is important to an extent that it should be included as part of the prediction outcome. Specifically, it has been implicitly conjectured in the Kaggle report that normalizing production by perforation length – which is the distance between upper and lower perforation depths – is beneficial for prediction. Therefore the variables describing different depths in the drill hole were investigated, and the effect of normalizing the first-12-months-of-production outcomes as suggested by Kaggle was explored.

There are (at least) seven variables in the IHS data set which describe different actual and projected depths of the wells: Perforation.Upper, Perforation.Lower, Depth.True.Vertical, Depth.Total.Maximum, Depth.Total.Driller, Depth.Total.Logger, Depth.Total.Projected.

The content of these variables is not equal, but they fall into three classes with similar content:
(1) Perforation.Upper, Depth.True.Vertical,
(2) Perforation.Lower, Depth.Total.Maximum, Depth.Total.Driller, Depth.Total.Logger, and
(3) Depth.Total.Projected.

Figure 4.2.1 shows density plots for one representative picked from each of the three classes. Upper and lower perforation is bimodal for gas wells and unimodal with a shoulder for oil wells. Total projected depth is tetramodal.

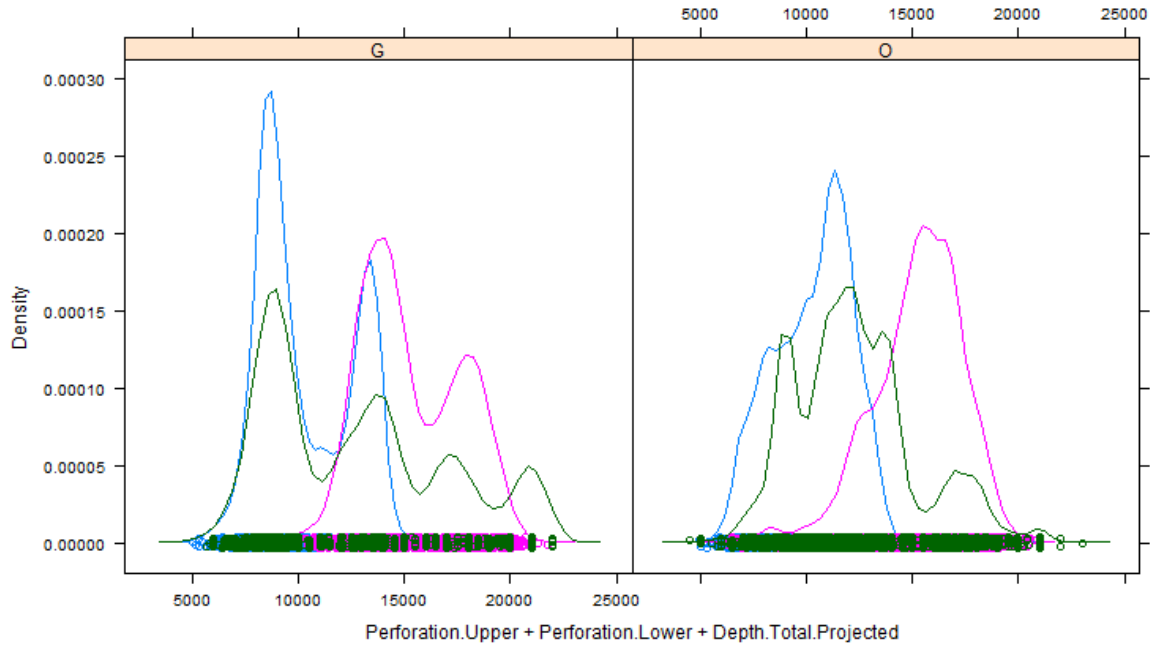

Figure 4.2.1: density plots of Perforation.Upper (blue), Perforation.Lower (violet) and Depth.Total.Projected (green), in feet, stratified by whether the primary product of the well is gas (left half) or oil (right half).

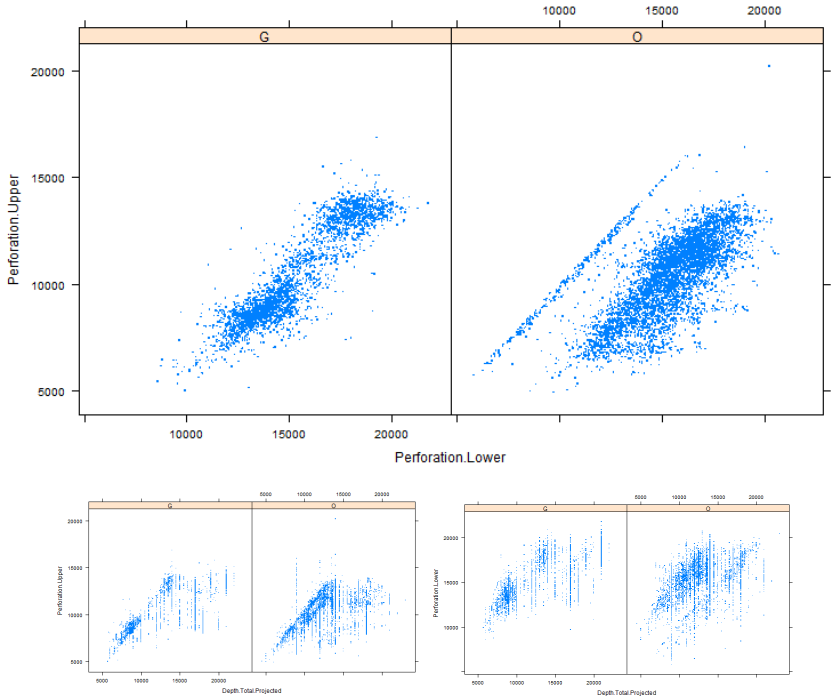

Figure 4.2.2: 2D scatter plots of Perforation.Upper, Perforation.Lower, Depth.Total.Projected, in feet, against each other. Stratified by whether the primary product of the well is gas (left panels) or oil (right panels).

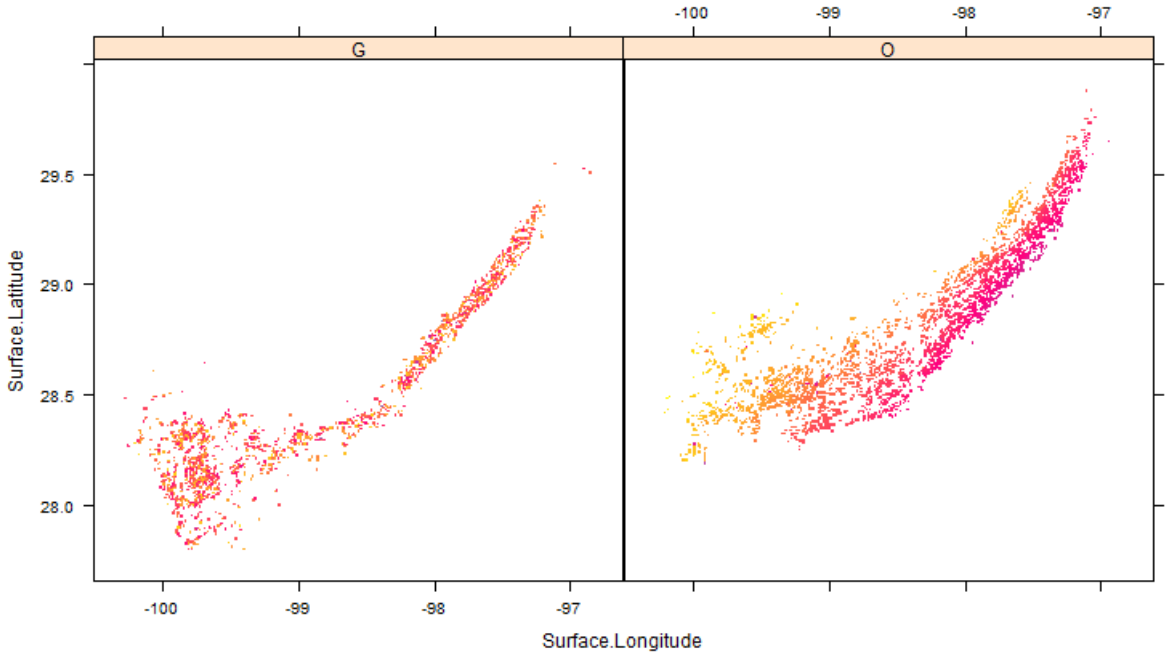

Figure 4.2.3: colored map plots of upper perforation depth. Red = high depth. Stratified by whether the primary product of the well is gas (left panel) or oil (right panel).

Figure 4.2.2 shows 2D scatter plots of one representative picked from each of the three classes against each other. Perforation.Upper vs Perforation.Lower exhibits interesting behaviour: for both oil and gas wells, Perforation.Lower is ca 5000 feet away from Perforation.Upper for most wells. For gas wells, there are further two clusters, while there are some oil wells where Perforation.Upper and Perforation. are almost equal.

Figure 4.2.3 shows an exemplary map plot of upper perforation depth (Perforation.Upper in the data set). It can be qualitatively observed that depth is geographically smoother for oil wells than for gas wells. The behaviour is similar for the other six variables. This is interesting when compared to the qualitatively reverse behaviour of oil production vs gas production. The plots involving Depth.Total.Projected feature discretization artefacts which could be explained by Depth.Total.Projected being rounded.

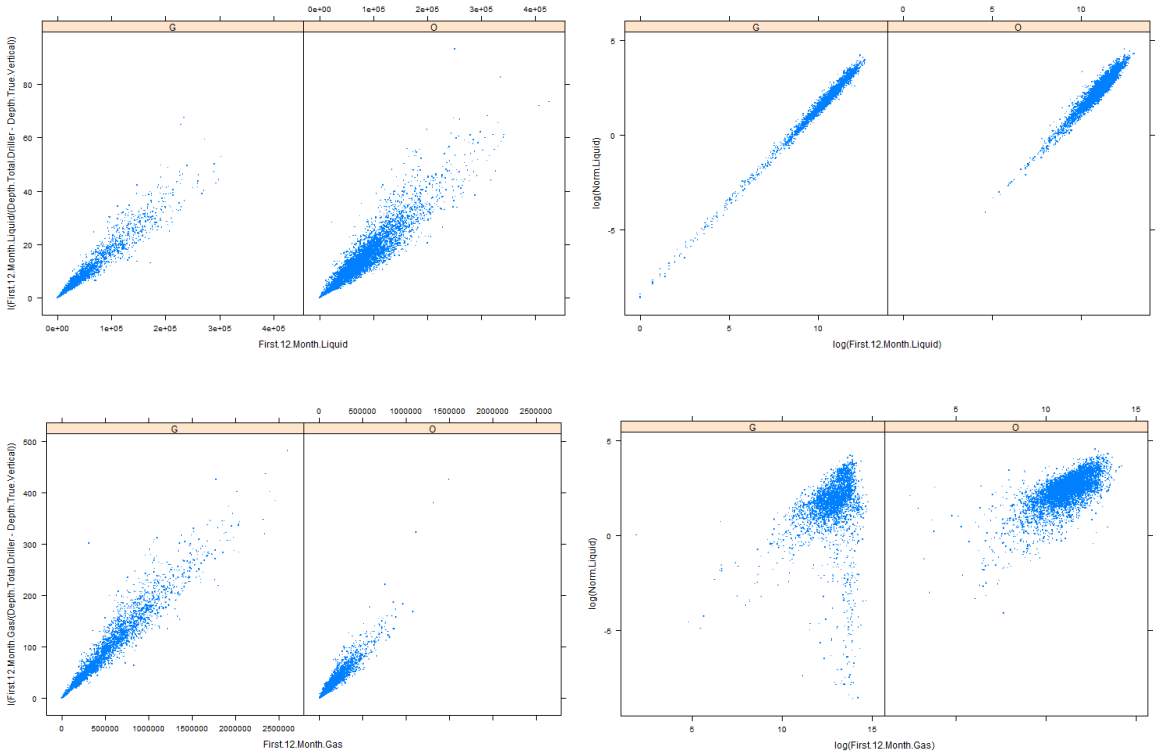

Figure 4.2.4: 2D scatter plots of first-12-months production/bbl versus first-12-months production/bbl divided by perforation length = Perforation.Lower minus Perforation.Upper in feet. Plotted for oil production (upper two panels) and gas production (lower two panels), plain production (left two panels) and logarithmic production (right two panels). Stratified by whether the primary product of the well is gas (left halves) or oil (right halves).

Figure 4.2.4 investigates the effect of normalizing first 12 months of production by perforation length = lower perforation depth minus upper perforation depth. The normalization appears to affect oil production less than gas production.

Summarizing preliminary conclusions:

(1) the depth variables exhibit interesting behaviour which may become relevant when predicting
(2) the analyses above do not offer insight on the effect of Kaggle's normalization by perforation depth; in particular, there is no exploratory evidence for a negative or positive effect

Whether these observations are indeed relevant in the prediction task can only be validated and quantified in the context of a complete predictive model.

Repository reference:
dataAnalysis/IHS - merged logs exploration/exploration_IHS_merged.R
dataAnalysis/IHS - merged logs exploration/IHSdata_master.R (master file)

4.3 SPECTRAL EXPLORATION OF PRODUCTION DATA

The original format of oil and gas production in the IHS data are monthly production curves, and the summaries in the tables of the data set are aggregates obtained from those curves. Thus, one natural way to understand the projection outcomes and approach question (A) is to study the production curves as curves.

One way of preserving the curve structure is putting the curves in a matrix where rows are wells and columns are different months of production, and then studying the spectral properties of this matrix. Since production is always a positive number, singular value decomposition was chosen as a method above principal component analysis (the main difference lies in the absence of row centering for the singular value decomposition).

Decline curve analysis estimates parametric summaries (e.g. slope and height) from single curves which are specified in advance, while a spectral decomposition can identify prototype curves without prescribing their shape.

Singular value decomposition was applied to the matrix of monthly production data described above, and the matrix of logarithmic monthly production data (see section 4.1, observation 3). Separate matrices were constructed for oil and gas production, yielding a total of four matrices of which singular value decomposition was computed.

In the matrices, production curves were shifted left so the first column contained only non-zero entries. Zeroes were filled in for missing and censored values (one may want to study more sophisticated treatment of censoring and missing values in future analyses).

Figures 4.2.1, 4.2.2, 4.2.4 show histograms and scatter plots for monthly oil production singular scores. Figures 4.2.6, 4.2.7, 4.2.9 show histograms and scatter plots for monthly gas production singular scores. Figures 4.2.3 and 4.2.8 show the corresponding singular vectors; the first singular vector can be interpreted as the principal prototype for a production curve, explaining most of the variation in the production curves. Second, third and further singular vectors are the principal lower-order corrections to the first singular vector.

In every case, for plain and logarithmic production, and oil and gas wells, three singular scores explain around 90% of in-sample variation. A very unusual and therefore potentially relevant finding is the spiral structure of the production curves when plotted by the log-singular scores, see figures 4.2.4 and 4.2.9. This indicates that the production curves may be later concisely summarized by one or more derived non-linear summaries - e.g. the position along the spiral.

In particular, since first-12-months of production is a summary which is linear in the production curves, that variable will not capture all information present in the production curves.

Figure 4.2.5 shows a scatter plot of the first logarithmic singular score vs cumulative oil production at the latest date.

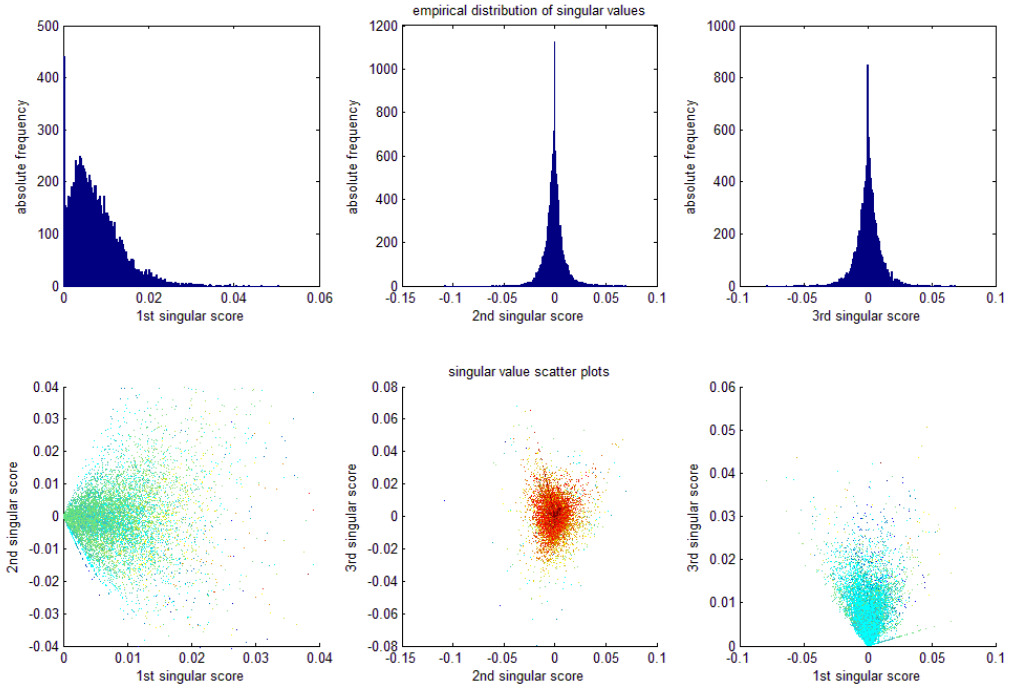

Figure 4.2.1: histograms and 2D scatter plots of first three singular scores for monthly oil production curves in the IHS production workbook/monthly production table. Colour is the third singular score not on x/y axes.

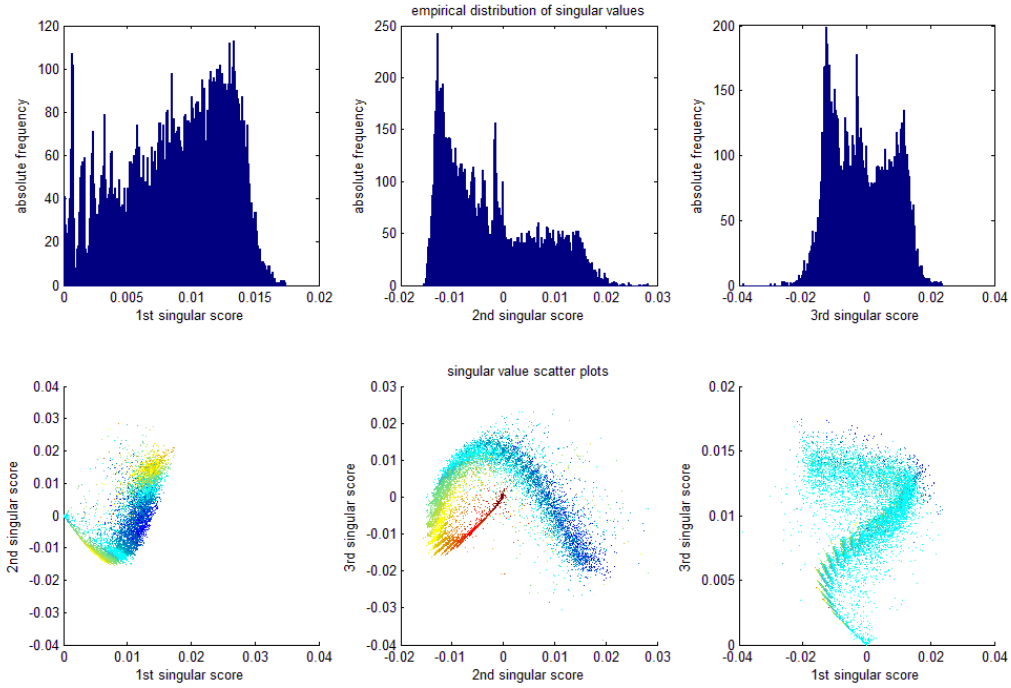

Figure 4.2.2: histograms and 2D scatter plots of first three singular scores for *logarithmic* monthly oil production curves in the IHS production workbook/monthly production table. Colour is the third singular score not on x/y axes.

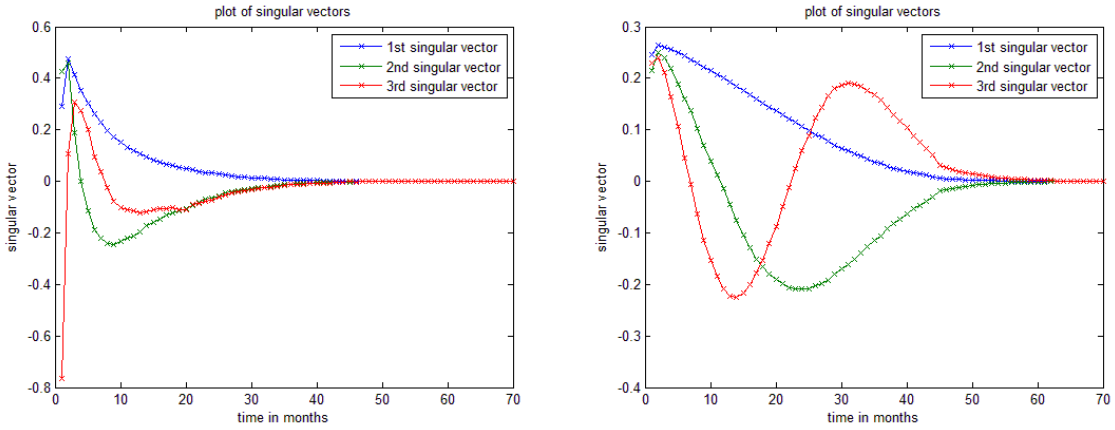

Figure 4.2.3: first three singular vectors of monthly oil production (left) and logarithmic oil production (right) in the IHS production workbook/monthly production table. x-axis = time in months after first production, y-axis = entry of singular vector. First non-logarithmic singular vector explains 78% of in-sample variance, first three explain 89%. First logarithmic singular vector explains 76% of in-sample variance, first three explain 91%.

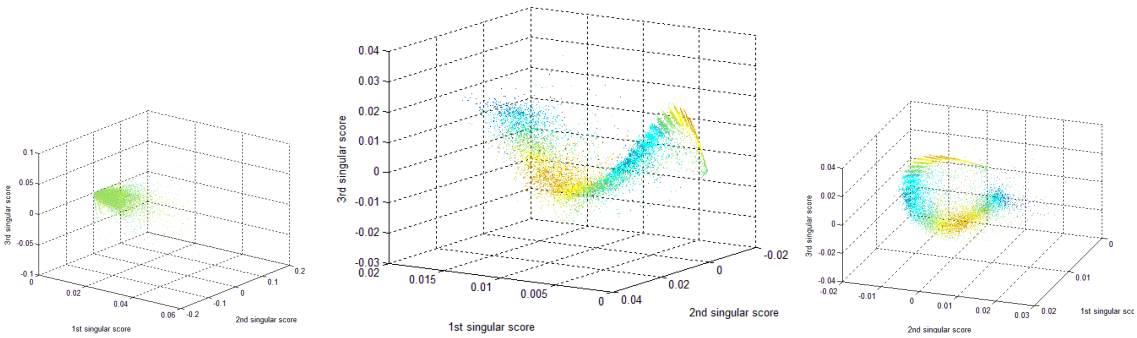

Figure 4.2.4: scatter plots of first four singular scores of monthly oil production (left) and logarithmic oil production (middle and right, different perspectives) in the IHS production workbook/monthly production table. First three singular scores are on the axis, colour is the fourth singular score.

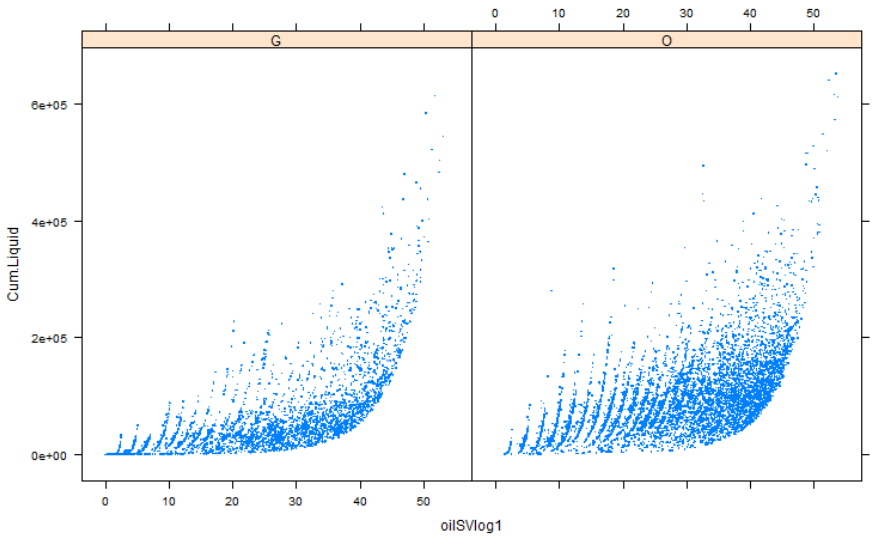

Figure 4.2.5: 2D scatter plots of first logarithmic singular score of monthly oil production vs cumulative oil production at most recent time of measurement. Stratified by whether the primary product of the well is gas (left half) or oil (right half).

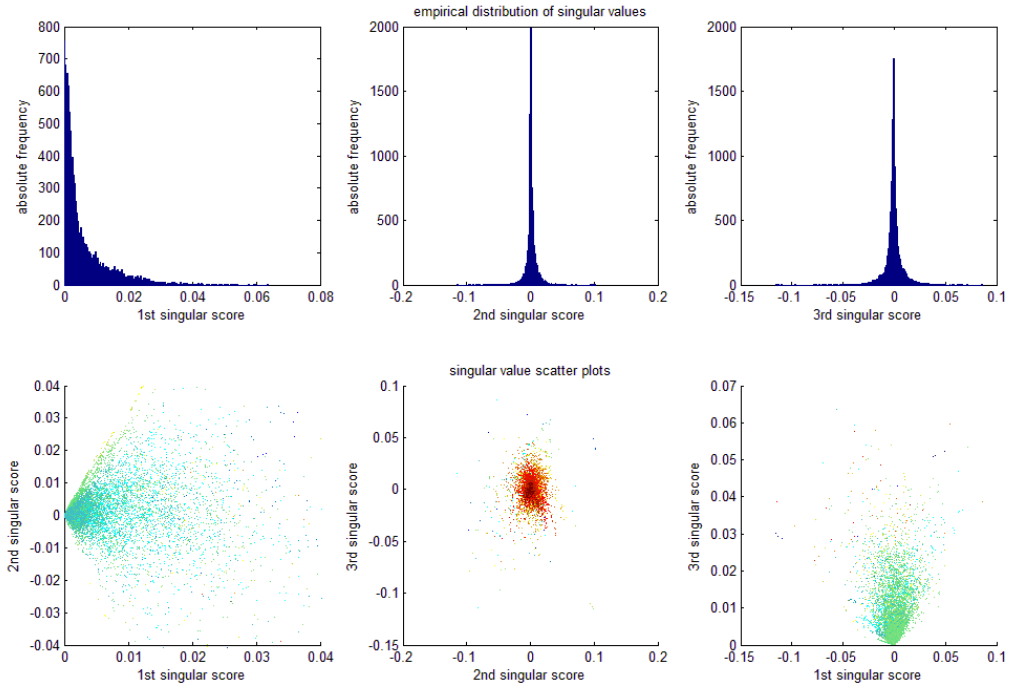

Figure 4.2.6: histograms and 2D scatter plots of first three singular scores for monthly gas production curves in the IHS production workbook/monthly production table. Colour is the third singular score not on x/y axes.

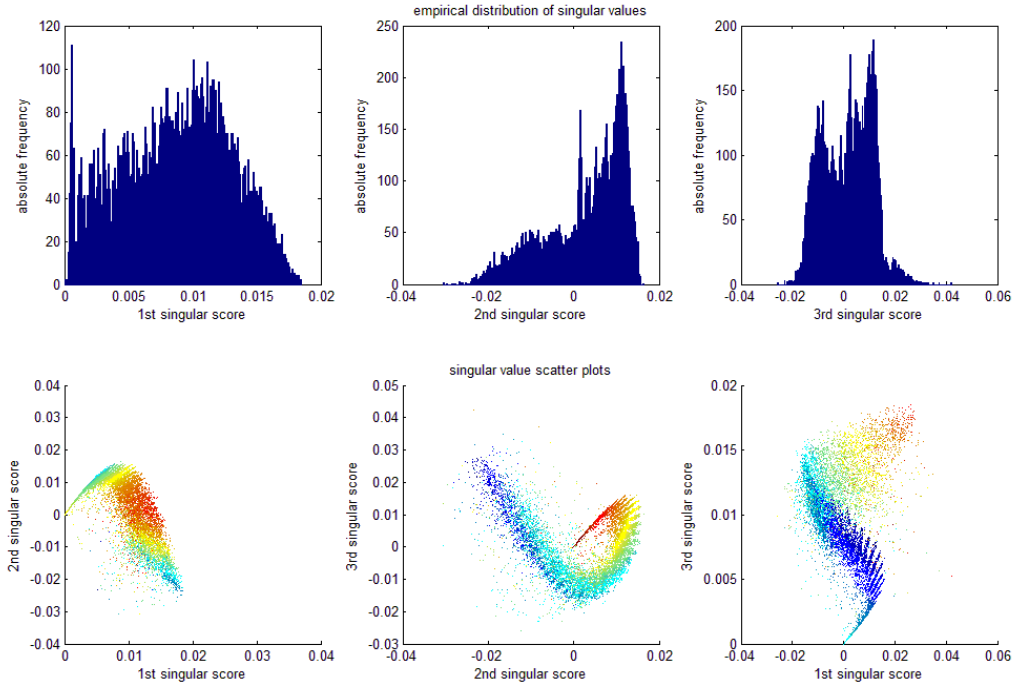

Figure 4.2.7: histograms and 2D scatter plots of first three singular scores for *logarithmic* monthly gas production curves in the IHS production workbook/monthly production table. Colour is the third singular score not on x/y axes.

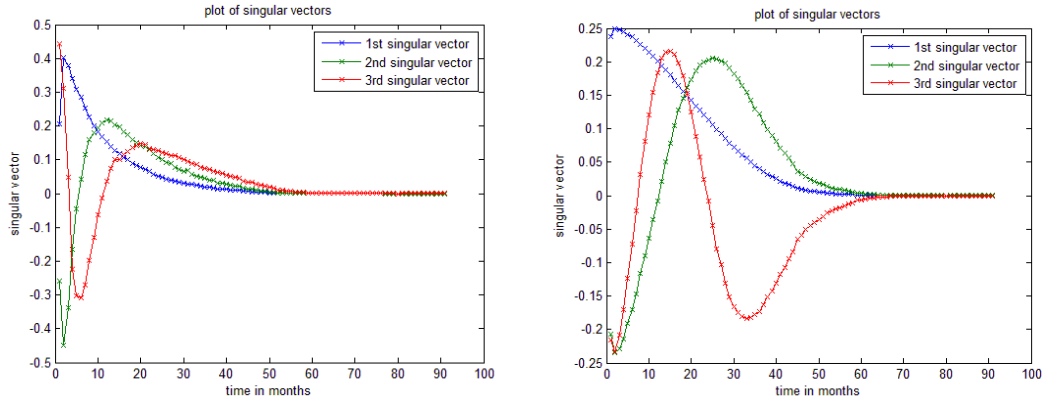

Figure 4.2.8: first three singular vectors of monthly oil production (left) and logarithmic oil production (right) in the IHS production workbook/monthly production table. x-axis = time in months after first production, y-axis = entry of singular vector. First non-logarithmic singular vector explains 78% of in-sample variance, first three explain 90%. First logarithmic singular vector explains 77% of in-sample variance, first three explain 90%.

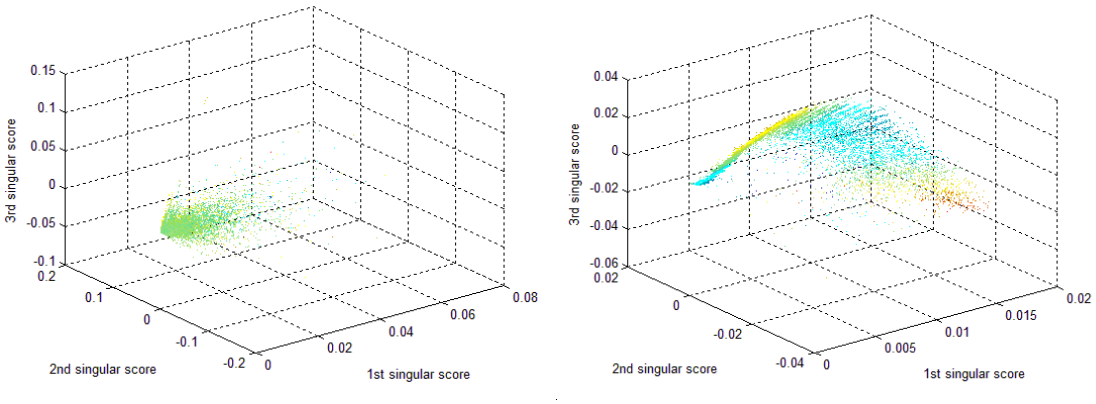

Figure 4.2.9: scatter plots of first four singular scores of monthly gas production (left) and logarithmic gas production (right) in the IHS production workbook/monthly production table. First three singular scores are on the axis, colour is the fourth singular score.

Concluding, the findings above indicate that:

(1) the shape of production curves can be concisely described by logarithmic singular scores
(2) the singular scores allow to expose hidden structure and may be therefore relevant in prediction
(3) the singular scores may prove crucial in finding a good proxy for EUR, or in estimating EUR

Whether these observations are indeed relevant in the prediction task can only be validated and quantified in the context of a complete predictive model.

Repository reference:

dataAnalysis/IHS - production - spectral analysis/exp_MP_svd.m
dataAnalysis/IHS - production - spectral analysis/exp_MP_spectral.m (master file)

5 ANALYSIS OF THE KAGGLE DATA SET

The Kaggle data set was second to be looked at in detail, as the natural starting point for question (D). Also, since it contained a pre-selection of covariates that were already interpolated, it offered an opportunity to study question (B) with quantitative methods by skipping the otherwise necessary interpolation step.

5.1 PERCENTILE RECOVERY CURVES

In their analyses, Kaggle have used the binary variable of being in the top quartile of first 12 months of production divided by perforation length as an outcome. As discussed in section 2.2 (A.iii), the choice of this outcome is non-standard, seemingly arbitrary, and not validated by comparison to alternatives.

A further problem is that since this measure of goodness is non-standard, it is lacking the theoretical statistical groundwork which usually allows to state whether a result is likely to be random or not – in particular it is difficult to decide whether a given method is significantly different from a random guess.

As one solution to this problem below, standard measures of prediction goodness such as the root mean squared error (RMSE) and the mean absolute error (MAE) of prediction will be reported.

In order to also have a measure of goodness that is comparable to and compatible with Kaggle's choice, but less arbitrary, "percentile recovery curves" were plotted. These are plots that given a prediction method, show for a percentile on the x-axis the out-of-sample fraction of values above the percentile correctly identified as being above – as an example, see figure 5.1.1 below as an example. A method can be represented by a curve, error bars (dotted) can be obtained by re-sampling. A random guess, in expectation, would achieve as percentile recovery curve the red straight connecting the points (0,0) and (100,100) drawn in every plot. A method that is better than random would swerve to the top left, with error bars optimally not touching that red baseline curve.

This graphical way of presenting and comparing methods alleviates some problems in Kaggle's choice of goodness measure, by adding a quantification of uncertainty/randomness, making comparison to a random guess and other baselines possible, and avoiding the arbitrary choice of the top quartile as a threshold. Analogue plots and recovery curves will be obtained for the identification of bottom percentiles.

It needs to be noted that even though the comparison via "percentile recovery curves" is less arbitrary than Kaggle's measure of goodness, it is still non-standard and appears to be novel (to the author of this report), thus is likely lacking supporting theory or may be problematic in unknown ways. Though similar concepts are known for ranking methods, such as Kendall's rank distance or, more generally, rank error measures.

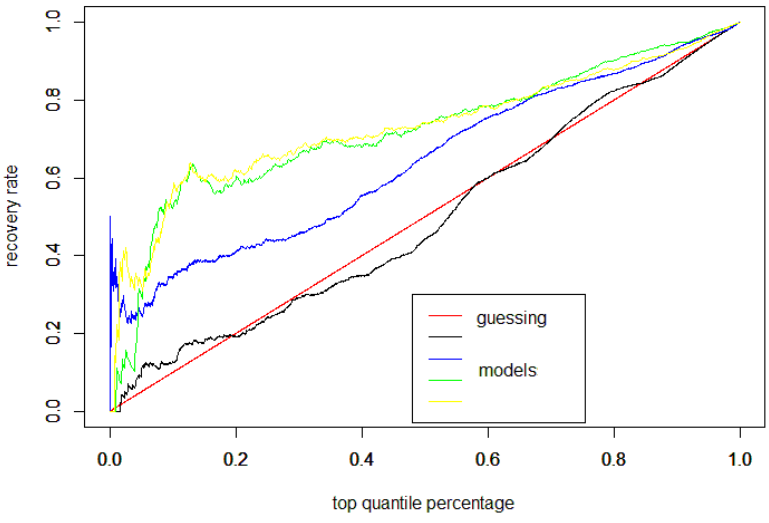

Figure 5.1.1: example plot for percentile recovery curves. Differently coloured curves correspond to different methods of prediction. The red straight is the curve which would be achieved in expectation by a random guess. The black prediction method is not much different from a random guess. A "good" predictor would swerve towards the top left, spanning the more area the better the method is – green and yellow methods are better than the blue one. As the curve approach the left, fluctuations are to be expected due to normalization effects.

5.2 PREDICTION FROM THE RULES VARIABLES

In order to determine whether the variables used by the Big Rules model are useful in predicting a prediction outcome, (untuned/default) off-the-shelf methods were employed to obtain predictions for first 12 months of oil production on the Kaggle rules data set in several validation setups. Measures of prediction goodness RMSE and MAE were obtained from the out-of-sample predictions, as well as percentile recovery curves.

The selected off-the-shelf methods were two versions of kernel support vector regression, as implemented in the kernlab package in R, and Breiman's random forests, as implemented in the randomForest package in R (see dataAnalysis/Kaggle data sets – exploration/CV-scripts/CV_load_offshelfML.R). No external parameter tuning was conducted. As "random guessing" baselines, prediction of the training mean and training median were considered. The methods were also compared to the predictions obtained from Kaggle's big rules and black box model, though the comparison may be flawed since it is not known whether the values obtained from Kaggle's black box are actually predictions, or if yes, whether they are all out-of-sample.

Two validation setups were considered: in the first, the out-of-sample error was estimated by standard five-fold cross-validation. In the second, 20 uniform folds were sampled, and 20 train/test splits were obtained where 5% of the data was used for *training* and 95% of the data was used for *testing*, in order to emulate an early state of the play.

Table 5.2.1 shows the quantitative results for RMSE and MAE, figures 5.2.2 and 5.2.3. show the quantile recovery curves. The off-shelf methods using the geological covariates present in the big rules data set make predictions of normalized production that are significantly better than random guesses.
It is also seen that Kaggle's black box model is easily outperformed by the off-shelf methods when the majority of the wells are available for training. When only 5% of the data is available for training, Kaggle's black box model is not significantly different from the off-shelf models, while it seems to be slightly worse in tendency. Again, it needs to be emphasized that it is unknown how much data was used in Kaggle's model. The percentile recovery plots imply that Kaggle's big rules model is, for most percentiles, not different from a random guess.
Further experiments were conducted that involved a geographical train/test split inspired by the rather vague description of such a validation procedure in Kaggle's report. The results do not differ qualitatively from the ones below (for the same numbers of training wells).

Conclusions from the above:

(1) (Kaggle's interpolates of) the covariates in the big rules model are useful in predicting oil production
(2) the performance of Kaggle's black box model (on oil wells) can be easily matched or exceeded by off-shelf methods, measured by both standard metrics and percentile recovery curves
(3) Kaggle's big rules model (on oil wells) is as good as a random guess in identifying top percentiles, while it appears to identify some of the *bottom* percentile wells better than a random guess.

number of training wells	*Mthd:* / *error msre.*	"random guessing"		off-shelf machine learning methods			Kaggle's models	
		training mean	training median	nu-kSVR	eps-kSVR	random forests	Big Rules model	Black box model
2104 wells (5-fold CV)	**RMSE**	9.8 ± 0.5	10.1 ± 0.5	6.9 ± 0.4	6.6 ± 0.3	5.6 ± 0.3	288 ± 2.6	8.6 ± 0.4
	MAE	7.4 ± 0.3	7.2 ± 0.3	5.2 ± 0.2	4.6 ± 0.2	3.9 ± 0.2	279 ± 3.3	6.7 ± 0.2
131 wells (rev 20-fold)	**RMSE**	9.8 ± 0.5	10.1 ± 0.5	8.1 ± 0.7	7.8 ± 0.8	7.7 ± 0.7	288 ± 5.1	8.6 ± 0.6
	MAE	7.4 ± 0.3	7.2 ± 0.3	6.2 ± 0.5	5.6 ± 0.5	5.7 ± 0.5	279 ± 6.6	6.6 ± 0.5

Table 5.2.1: out-of-sample RMSE and MAE of predictions obtained from selected methods on the rules data set in two validation setups. Standard errors are batch means of standard errors on the test set.

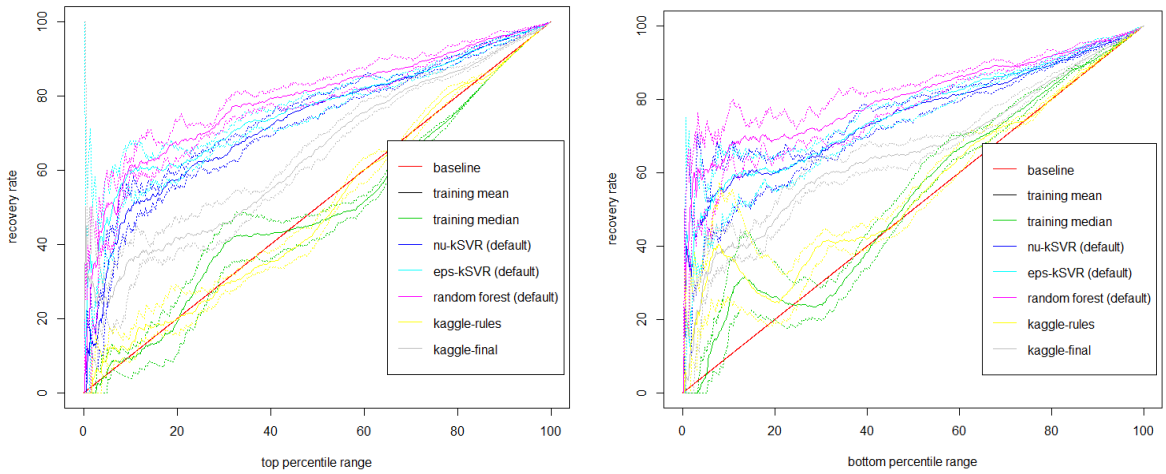

Figure 5.2.2: top and bottom percentile recovery curves for Kaggle's two models (the big rules-based "kaggle-rules" and the black box model "kaggle-final") and various naïve baselines predicting Kaggle's normalized production outcome from the interpolated geological variables in the rules data set. Solid curves are the median recovery over the five train/test splits of 5-fold cross-validation, dotted curves are min/max recovery over the same folds. Curve for training mean is the same as for training median.

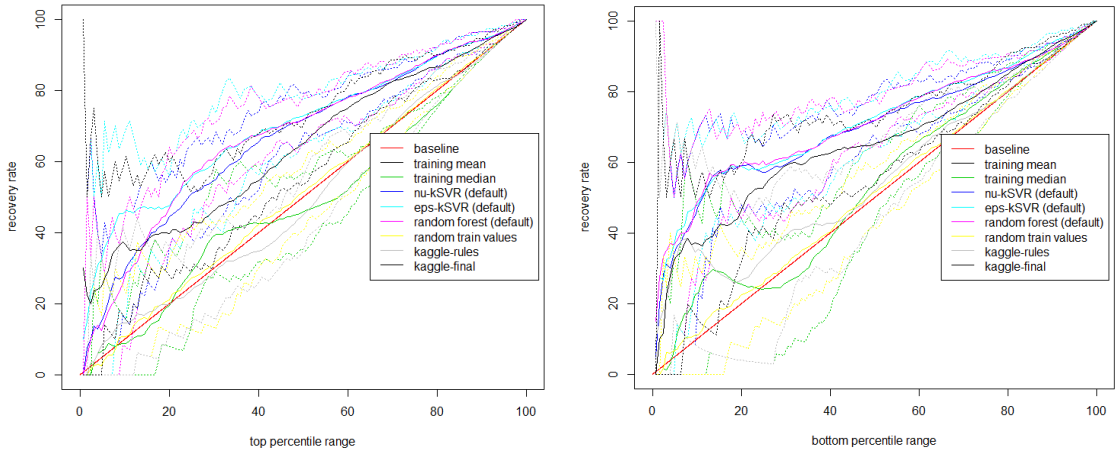

Figure 5.2.3: top and bottom percentile recovery curves for Kaggle's two models (the big rules-based "kaggle-rules" and the black box model "kaggle-final") and various naïve baselines predicting Kaggle's normalized production outcome from the interpolated geological variables in the rules data set. Solid curves are the median recovery over the twenty train/test splits of reversed 20-fold cross-validation (= 5% train/ 131 wells and 95% test splits), dotted curves are min/max recovery over the same folds. Curve for training mean is the same as for training median.

Repository reference:

dataAnalysis/Kaggle data sets – exploration/experiment_RuleFeatures_pred.R

dataAnalysis/Kaggle data sets – exploration/master_Kaggle_explore.R (master file)

5.3 PREDICTION FROM THE BLACK BOX MODEL VARIABLES

The analogue of the experiment in section 5.2 was conducted for the interpolated/integrated variables present in the Kaggle black box data set.

As one additional prediction method not present in 5.2, ordinary least squares regression was considered. Furthermore, for each of the two validation set-ups (5-fold and reverse 20-fold cross-validation) the methods were instantiated on three different sets of variables: (1) position covariates (longitude/latitude) only, (2) the interpolated geological covariates in the black box data set, but not the position covariates, and (3) both position and geological covariates.

Table 5.3.1 shows the results for RMSE and MAE; percentile recovery curves for 5-fold cross-validation are shown in figure 5.3.3. Curves for reverse 20-fold look qualitatively similar to figure 5.2.3.

validation set-up	*covariates used for prediction*	*Mthd: error msre.*	"random guessing" training mean	"random guessing" training median	least-squares regression	off-shelf machine learning methods nu-kSVR	off-shelf machine learning methods eps-kSVR	off-shelf machine learning methods random forests
5-fold CV 80% = 2104 training wells	*(1) position (lon-lat/N-S)*	**RMSE**	10 ± 0.5	10 ± 0.5	8.6 ± 0.4	7.3 ± 0.4	7.1 ± 0.4	5.7 ± 0.3
		MAE	7.6 ± 0.3	7.4 ±0.3	6.3 ± 0.3	5.5 ± 0.2	4.9 ± 0.2	3.9 ± 0.2
	(2) geology (Kriged vars)	**RMSE**	10 ± 0.5	10 ± 0.5	7.3 ± 0.4	6.7 ± 0.3	6.4 ± 0.4	5.1 ± 0.3
		MAE	7.6 ± 0.3	7.4 ±0.3	5.4 ± 0.2	5.2 ± 0.2	4.4 ± 0.2	3.5 ± 0.2
	(3) both of the above	**RMSE**	10 ± 0.5	10 ± 0.5	7.3 ± 0.4	6.7 ± 0.3	6.4 ± 0.4	5.2 ± 0.3
		MAE	7.6 ± 0.3	7.4 ±0.3	5.3 ± 0.2	5.1 ± 0.2	4.4 ± 0.2	3.5 ± 0.2
reverse 20-fold CV 5% = 131 training wells	*(1) position (lon-lat/N-S)*	**RMSE**	10 ± 0.9	10 ± 1.0	8.2 ± 0.8	8.2 ± 0.7	8.1 ± 0.8	8.0 ± 0.9
		MAE	7.7 ± 0.6	7.4 ± 0.6	6.5 ± 0.5	6.3 ± 0.5	5.7 ± 0.5	5.8 ± 0.5
	(2) geology (Kriged vars)	**RMSE**	10 ± 0.9	10 ± 1.0	8.9 ± 0.8	8.0 ± 0.7	7.8 ± 0.7	7.6 ± 0.7
		MAE	7.7 ± 0.6	7.4 ± 0.6	6.5 ± 0.5	6.1 ± 0.4	5.7 ± 0.5	5.4 ± 0.5
	(3) both of the above	**RMSE**	10 ± 0.9	10 ± 1.0	8.9 ± 0.8	7.9 ± 0.7	7.8 ± 0.7	7.5 ± 0.7
		MAE	7.7 ± 0.6	7.4 ± 0.6	6.5 ± 0.5	6.1 ± 0.4	5.7 ± 0.5	5.4 ± 0.5

Table 5.3.1: out-of-sample RMSE and MAE for predictions of normalized oil production, obtained from selected methods on the black box data set in two validation set-ups for three choices of covariates. Standard errors are batch means of standard errors on the test set. The Kaggle predictions are always the same, since the Kaggle model cannot be trained. Direct comparison in the validation settings is therefore impossible.

Mthd: error msre.	Kaggle's black box model
RMSE	8.7 ± 0.4
MAE	6.8 ± 0.2

validation set-up	*covariate that is predicted*	*Mthd: error msre.*	"random guessing" training mean	"random guessing" training median	least-squares regression	off-shelf machine learning methods nu-kSVR	off-shelf machine learning methods eps-kSVR	off-shelf machine learning methods random forests
5-fold CV 80% = 2104 training wells	*Longitude /0.01°*	**RMSE**	91 ± 2.1	92 ± 1.7	2.5 ± 0.3	5.7 ± 1.6	7.9 ± 1.4	2.1 ± 0.5
		MAE	82 ± 1.7	81 ± 1.8	1.6 ± 0.1	0.7 ± 0.2	4.4 ± 0.2	0.5 ± 0.1
	Latitude /0.01°	**RMSE**	41 ± 1.3	41 ± 1.4	2.0 ± 0.2	5.1 ± 1.6	5.6 ± 1.5	2.4 ± 0.1
		MAE	35 ± 0.1	35 ± 0.1	1.2 ± 0.1	0.5 ± 0.2	2.1 ± 0.2	0.4 ± 0.1
reverse 20-fold CV 5% = 131 training wells	*Longitude /0.01°*	**RMSE**	91 ± 3.1	91 ± 3.1	6.0 ± 1.8	19 ± 6.0	17 ± 4.5	11.5 ± 3.2
		MAE	82 ± 3.5	82 ± 3.5	1.9 ± 0.4	7.8 ± 1.4	8.8 ± 1.1	4.4 ± 0.8
	Latitude /0.01°	**RMSE**	41 ± 2.4	42 ± 2.7	8.9 ± 2.8	11.6 ± 3.9	11.9 ± 3.8	10.2 ± 3.3
		MAE	35 ± 1.9	35 ± 2.1	1.6 ± 0.6	2.8 ± 0.8	4.0 ± 0.8	2.5 ± 0.7

Table 5.3.2: out-of-sample RMSE and MAE for predictions of longitude/latitude from the geological covariates in the black box data set, employing selected methods in two validation set-ups. Standard errors are batch means of standard errors on the test set.

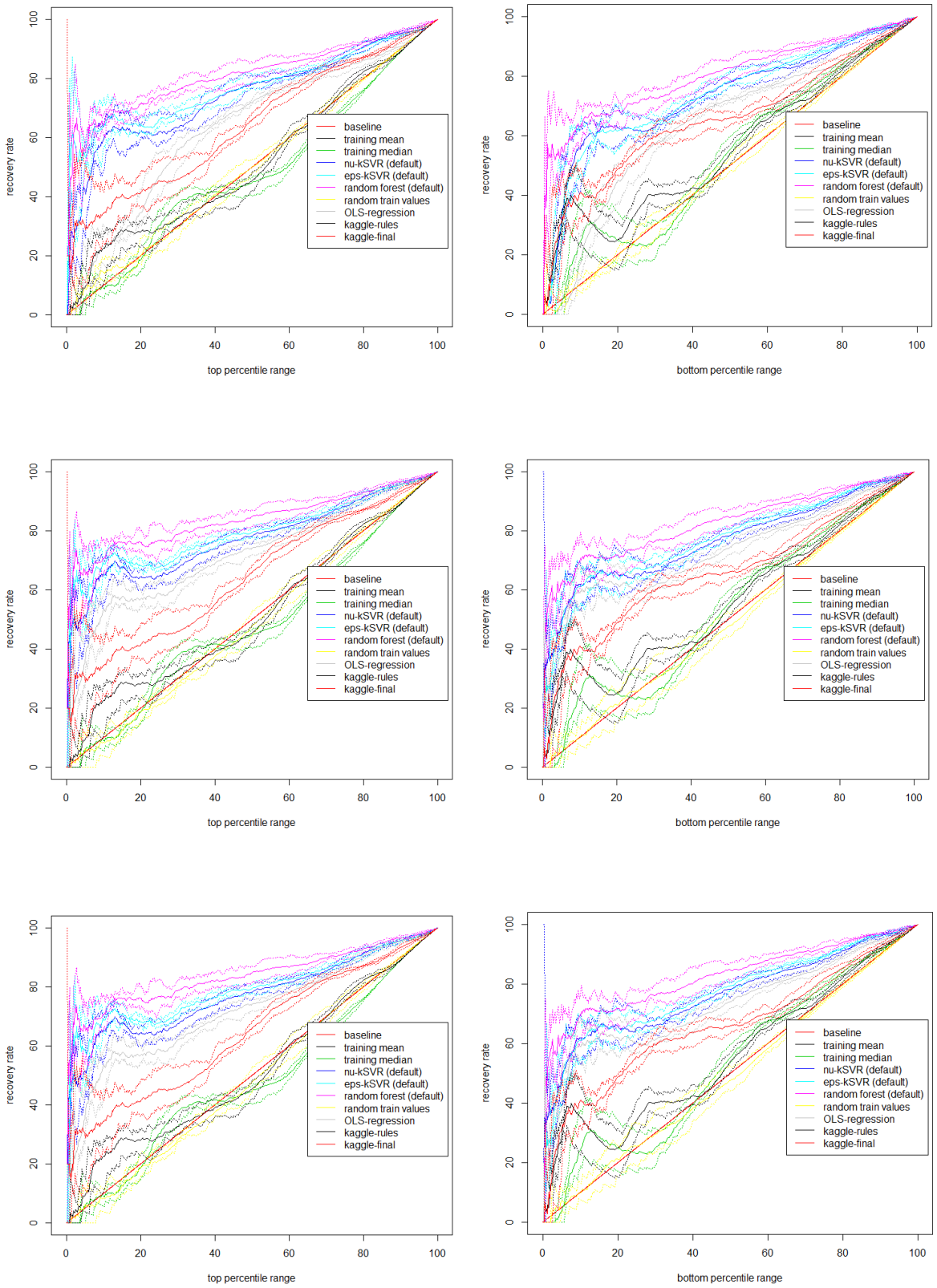

Figure 5.3.3: percentile recovery curves for predicting normalized oil production in the 5-fold cross-validation (80% = 2104 training wells set-up) from different sets of variables on the black box data set. Top row = (1) from position only; middle row = (2) from geological covariates only; bottom row = (3) from all covariates. Non-straight red and black curves are Kaggle's models; curve for training mean is identical to training median.

Generally, most methods can predict better in both setups (2) geological covariates and (3) geological and position covariates than when given only (1) position covariates. The only notable exception is least squares regression which becomes worse on 131 training wells when adding the geological covariates. Errors and curves are very similar for predictions from (2) geological covariates only and (3) geological and position covariates.

To understand the similarity between (2) and (3), it was investigated how well position covariates can be predicted from the geological covariates in the same validation setup. Table 5.3.2 shows errors of predicting position from geological covariates. All (non-guessing) methods can reconstruct latitude and longitude with high accuracy when given only the geological covariates. Here, among all methods, linear regression performs best, possibly because the selected machine learning methods are better at intrapolating, while regression is better extrapolating. The differences between MAE and RMSE indicate that most well positions can be reconstructed well, while there may be a few wells that are reconstructed very badly.

Conclusions from the above:

(1) comparatively good predictions can be already made by pure interpolation of (Kaggle's normalized) oil production.
(2) these predictions can be improved, on average, by adding (Kaggle's interpolates of) geological features – depending on the amount of training wells and the method
(3) adding geological features may damage a method's performance
(4) well location, in terms of longitude/latitude, can be almost perfectly (average error in the order of kilometres) reconstructed from the interpolated geological features; this may explain why quality of prediction remains unchanged when adding position above geological features
(5) Kaggle's black box model is slightly worse than pure interpolation of oil production from 131 nearby wells – where the latter method does not require geological variables to make a prediction

Repository reference:
dataAnalysis/Kaggle data sets – exploration/experiment_Kaggle_Kriging_pred.R
dataAnalysis/Kaggle data sets – exploration/master_Kaggle_explore.R (master file)

5.4 INFLUENCE OF NUMBER OF TRAINING PRODUCTION WELLS

Since in 5.2 and 5.3 goodness of prediction varied with (i) the size of the training set (bigger was better) and (ii) whether geological variables and/or position (longitude/latitude) were given to the learning method or not, a series of experiments was conducted on the interplay of these two parameters.

In the experiment, (i) size of the training size was varied between 10 and 2000 production wells for training, and (ii) either only position (longitude/latitude) was passed to the method, or position plus the geological variables in the black box data set. (Geological variables without position variables is not shown below, but as in 5.3 are qualitatively equivalent to geological variables plus position variables.)

Four off-the-shelf methods, least squares linear regression, two variants of support vector regression (as in 5.2), and Breiman's random forests (as in 5.2) were employed to predict normalized first 12 months of oil production.

The out-of-sample MAE was estimated on a test sample of 500 production wells, independently re-sampled in 10 repetitions of the validation setup.

Figure 5.4.1 shows the MAE of the four methods in dependence of (i) training sample size and (ii) whether geological variables are passed to the methods or not. The prediction error decreases with size of training sample. For comparison, the MAE of Kaggle's black box model is 6.6 ± 0.2 (see table 5.2.1), which is on average outperformed by a random forest using geological variables on 50 training wells, or a random forest using only position (longitude/latitude) on 100 training wells, and significantly by a random forest using geological variables on 150 training wells, or a random forest using only position (longitude/latitude) on 200 training wells.

Figure 5.4.2 shows, in a single plot for comparison, the average MAEs of the four methods. The plot indicates that in the presence of only few wells adding geological variables in the prediction may be detrimental, while the geological variables help when some bot not very many variables are available – numbers depending on the method. As more wells become available, it matters less and less whether geological features are added or not. Among the off-shelf methods, random forests are able to profit earliest from the geological features.

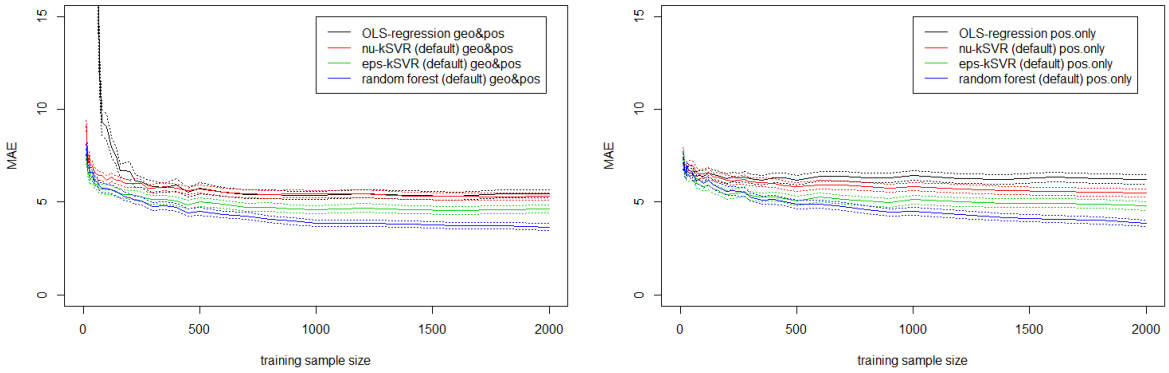

Figure 5.4.1: out-of-sample MAE of predicting Kaggle's normalized production from different sets of variables in the Kaggle black box data set. The test error was estimated by 10 repetitions of sub-sampling cross-validation with a test set of size 500 and a training set of size as specified on the x-axis. The standard error (dotted lines) is estimated from the sample of 10 repetitions. The variables used in the left panel are geology and position (longitude/latitude); in the right panel, predictions are made from position (latitude/longitude) only.

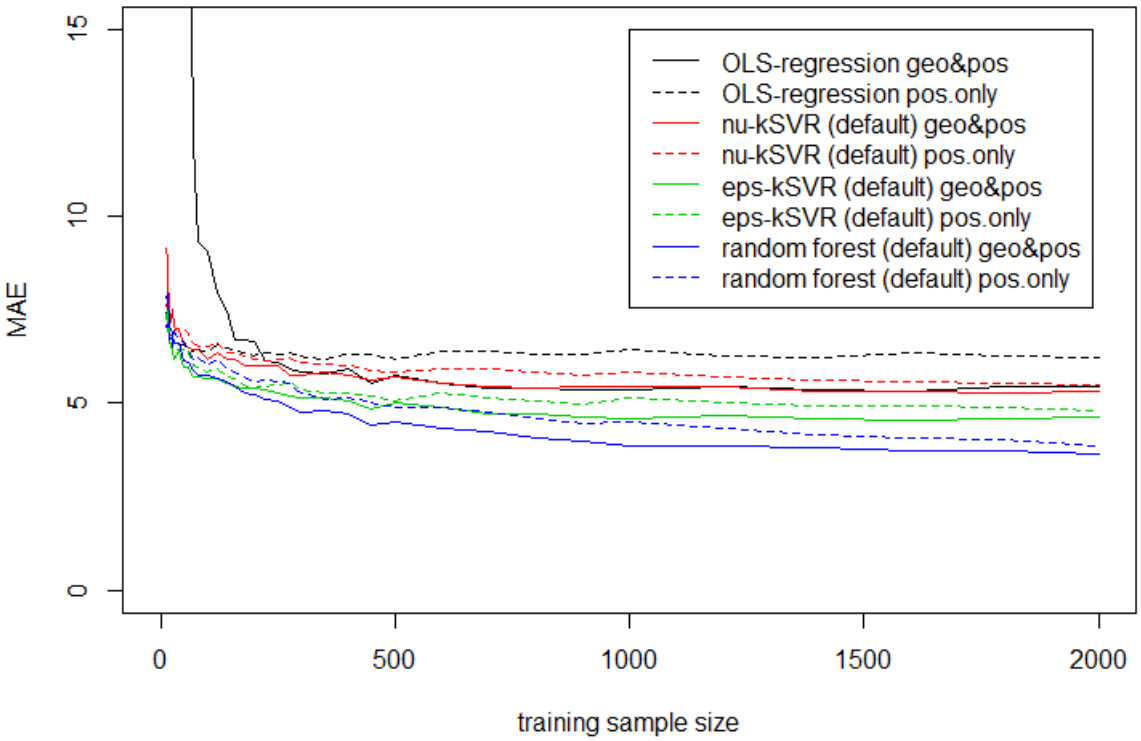

Figure 5.4.2: the means from both panels in figure 5.4.1, plotted for comparison in a single plot without error bars. Solid lines are mean prediction errors for models using both position (longitude/latitude) and geological covarates, dotted lines are mean prediction errors for models using position (longitude/latitude) only for prediction

Conclusions which may be drawn from the above:

(1) it is indicated that there are different "density regimes" for prediction: depending on the method and number of wells available, the prediction may profit – or not – from adding (all or a subset of) the geological features in the Kaggle black box data set
(2) one may conjecture that the more a method is considered "advanced", the earlier it may profit from addition of geological features, while adding in more variables may hurt more basic methods
(3) in terms of MAE, Kaggle's black box model can be outperformed (in predicting Kaggle's normalized oil production) by off-shelf methods on random training samples with as few as 100 production wells

The findings above will need to be checked again on interpolated/integrated variables obtained from the Core Labs data set, as it is unknown how the variables in the Kaggle data sets were obtained.

Repository reference:
dataAnalysis/Kaggle data sets – exploration/experiment_RuleFeatures_pred.R
dataAnalysis/Kaggle data sets – exploration/master_Kaggle_explore.R (master file)