NTNU

Norwegian University of
Science and Technology

Verification for Chance of Geologic Success

Chen Bao

Petroleum Geoscience and Engineering
Submission date: January 2016
Supervisor: Reidar B Bratvold, IPT

Norwegian University of Science and Technology
Department of Petroleum Engineering and Applied Geophysics

NTNU

Verification for Chance of Geologic Success

Chen Bao

January 2016

MASTER THESIS

Department of Petroleum Engineering and Applied Geophysics

Norwegian University of Science and Technology

Supervisor: Reidar B Bratvold

Preface

This master thesis was written during the autumn semester 2015 at Petroleum Engineering and Applied Geophysics Department, Faculty of Engineering, Science and Technology, Norwegian University of Science and Technology. The thesis is part of the five-year master's program: Petroleum Engineering and Petroleum Geosciences.

The topic was suggested by professor Reidar B Bratvold. I would like to express my gratitude to my supervisor, professor Reidar B Bratvold for his excellent guidance and support throughout my master thesis and master project.

Trondheim, 2016-01-24

Chen Bao

Summary

Prospect is an identifiable possible trap potentially containing petroleum. Chance of geologic success, i.e, probability of mobile hydrocarbon for a prospect is one key input for further evaluations of prospects. Chance of geologic success is commonly obtained by multiplying probability of essential geologic factors. In order for the subsurface hydrocarbon accumulation to exist, essential geologic factors must coincide. One example of essential geologic factors includes a reservoir rock, a trap, a source rock and a migration route. The probability of geologic factors are usually estimated by a group of experts and are largely subjective. Tools as risk table for evaluating probability of geologic factor provides correspondence between probability value and qualitative description, which helps to make consistent assessment. Experts are still susceptible to individual bias and group bias which influence the probabilities they generate. Different elicitation methods that helps to formulate a person's knowledge and beliefs about uncertain events into probabilities are shortly presented. Elicitations mitigates the effects of bias, but estimating reliable probabilities remains to be difficult.

Meteorologists, on the other hand makes comparatively reliable predictions. One reason is that forecast verification is extensively used in meteorology, so that their forecast is systematically studied along with actual observations. Post-drill analysis is increasingly used to improve the quality of predictions and estimations made by explorers, however statistical methods being used are very limited. Extensive forecast verification measures can provides multi-facets evaluation of probability prediction performance, leading to a more reliable probability prediction in the end.

Summary measures provides single scores for the overall quality of prediction performance. Distribution- oriented measures, based on joint, marginal and conditional distributions of predictions and results, provide detailed information about prediction quality from different angles in terms of various verification measures/attributes. Graphical measures including sharpness histogram, reliability diagram and discrimination diagram together provide a more complete picture of the forecast quality. ROC analysis remains exploratory for chance of geologic success and has the potential to support decision-making in exploration. Statistical methods as logistic regression and kernel density method which helps to improve estimates for verification

measures for small sized data is utilized and presented. All these measures are applied to real dataset of chance of geologic success and results. The strength and weakness of those probability assessment in specific probability interval are demonstrated and discussed for the real datasets.

Contents

Chapter 1

Introduction

1.1 Background

A prospect is an individual geological unit, which potentially contains accumulation of hydrocarbons. Prospect is also the basic unit where exploration and production decisions are made. A definition of prospect connecting exploration and operation activities is given by Harbaugh et al. (1995): "a specific locality within an area where we possess or may acquire a lease or concession and which we interpret to have geological or economic characteristic that may warrant testing by drilling". There are many stages in the life of an oil or gas field, from assessment of prospect, to development and production of the field. Prospect evaluation is the first stage in this process and is the one where the least amount of information is available and therefore the greatest uncertainty exists, and great financial and development decisions will be based on probability of occurrence of movable hydrocarbons (geologic success) and the volumes of hydrocarbons in a prospect are two of the most important tasks in prospect evaluation. The existence of movable hydrocarbon in a prospect is the base for all other analysis and decisions. How chance of geologic success is obtained will be discussed in this paper.

Poor performances such as high dry hole rate and over-optimism in high risk-explorations result from poor prediction performance are pointed out by Rose (2001). Humans are susceptible to bias caused by external factors and internal cognition and motivations. Explorers are non-exceptions. The pre-drill probability predictions based on geologists knowledge or belief are normally deemed as subjective where biases exist. How bias are diffused from explorers to

numerical probabilities will also be shortly discussed.

However, subjective probability assessment can be consistent and be improved. One common practice of post drill analysis is comparing outcomes and predictions of prospect.. But often only basic measures as simple statistical average are used in the analysis. Thus, very limited information and possible wrong interpretation is produced by the analysis, leading to no improvement in probability calibration. On the other hand, useful statistical procedures has been extensively explored and developed in forecast verification discipline, which can provide detailed information about strength and weakness of the forecasts, leading to an improved geologic success in the end. Verification measures will be applied to 3 sets of data of chance of geologic and result, to see what kind of problems can be detected.

Task of improving probability assessment in exploration can be tackled from different angles, such as setting up detailed and consistent protocols and standards, or applying elicitation methods, which is not a focus of this paper. Even for verification measures, the measures presented in this paper may not fit well the geologic probability assessments, due to my personal very limited knowledge and experiences of exploration process. These measures, nevertheless, provide statistically fit methods and exploratory suggestions to study the prognoses probabilities and their correspondence with the observations.

1.2 Objectives

The main objectives of this Master's project are:

- General review of probability prognosis of geologic success and of inconsistency/bias leading to unsatisfactory performance
- Introduce suitable verification measures for chance of geologic success prognosis
- Demonstrate verification measures' usefulness in assessing the quality of chance of geologic success assessment and in improving the calibration in the end.
- More objectives

The ultimate goal for introducing verification measures is to help explores to make improved probability assessments.

1.3 Datasets

The prognosis of chance of geologic success data and corresponding results data are not easily available, not only because E&P companies like to keep them private, but also because likely many E&P companies don't have well documented data of them. Three sets of data are found and are applied by verification measures.

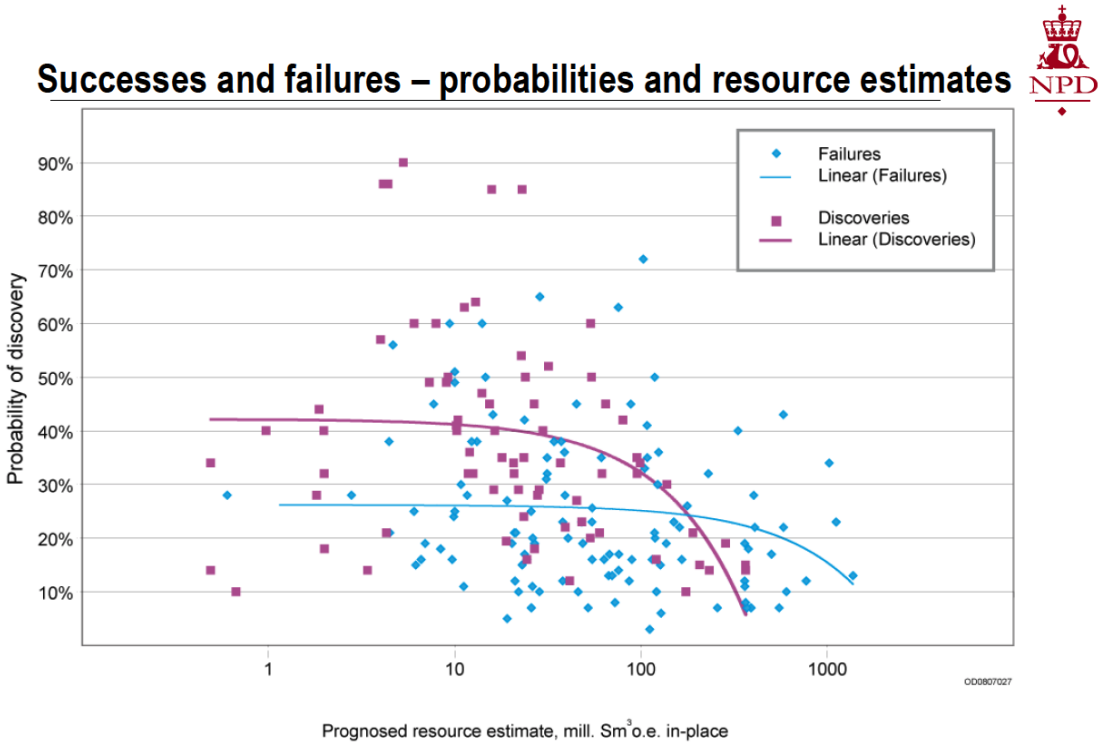


Figure 1.1: 184 pairs of probability of discovery and result (1998-2000) on NCS (Kari Ofstad and Helliksen, 2015)

1. Probability of discovery before drilling and observation data after drilling of 184 prospects on Norwegian continental shelf(NCS) from 1998 to 2007 from the Norwegian Petroleum Directorate (NPD). According to the Resource Management Regulations, the operators of wildcat wells are required to submit both the prognoses and results of wildcat wells to NPD. The data are extracted from graph 1.1 from NPDs presentation *Prognoses and results of wildcat wells drilled between 1998 and 2007 on the Norwegian Continental Shelf* (Kari Ofstad and Helliksen, 2015). The probability of discovery data are manually read from the graph, so there may be some small errors. 184 paris of prediction and observation data are obtained from the graph.

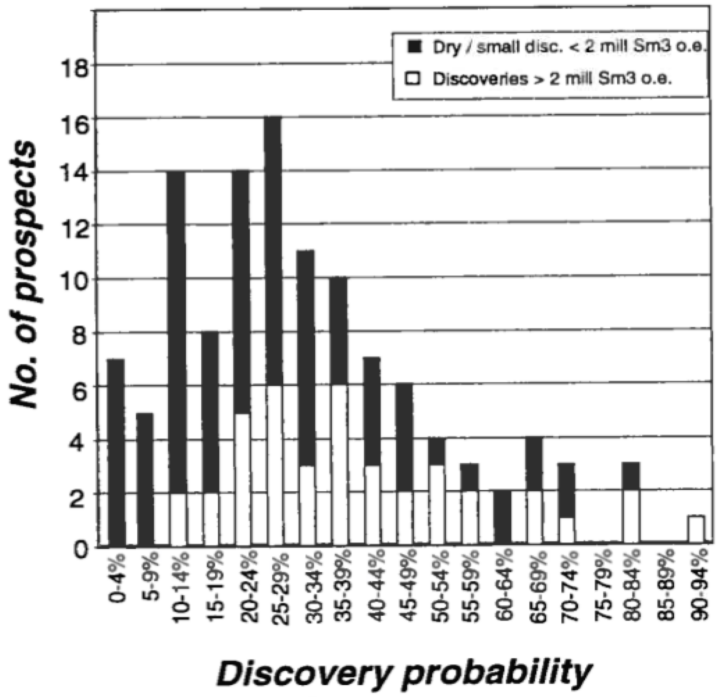
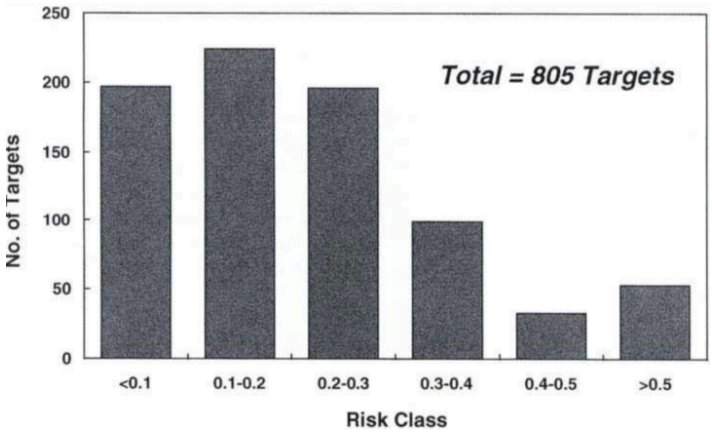


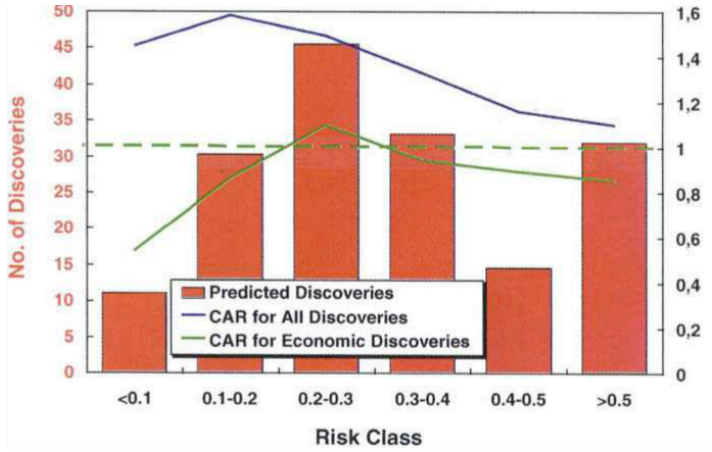
Figure 1.2: Discovery probability and result of 118 prospects (1990-1997) on NCS from NPD project (Ofstad et al., 2000)

2. Binned data of predicted discovery probability and observation of 118 prospects on Norwegian continental shelf (1990-1997) from NPD's evaluation of the 8th-14th licensing rounds. The data are extracted from the graph 1.2 from *Evaluation of Norwegian Wildcat Wells* (Ofstad et al., 2000). For each well one or several prospects are reported. so there could be several prognoses for one well, but only one result. As seen in the graph, original data are not available; but only categorized/binned prediction data are obtained.

3. Binned data of prediction and result of BP's 805 drilling targets drawn from over 40 countries from 1983 to 1997. The predicted chance of success are in 6 categories as shown in the graph 1.3a, The number of success targets and dry targets are acquired by calculating data read from the graph 1.3b. The graphs are from *Prediction accuracy in petroleum prospect assessment: A 15 year retrospective in BP* (Harper, 2000).



(a) Number of targets by predicted chance factor



(b) Predicted discoveries and accuracy ratio by chance factor

Figure 1.3: Graph for data of BP's 805 drilling targets (1983-1997) (Harper, 2000)

1.4 Structure of the Report

The rest of the report is organized as follows. Chapter 2 is a literature review of current evaluations of chance of geologic success. Chapter 3 discusses how bias influence explorers' probability assignments; then a short review of of elicitation theories which help to reduce and avoid bias; verification is introduced lastly and the benefits of verification is also presented. Chapter 4 are scalar measures which are also key attributes and concepts for forecast verification. Chapter 5 presents graphical verification measures by which users can acquire more direct information of the probability prediction performance and ROC analysis which is a more exploratory and experimental method for probability of geologic success. All these measures are applied to datasets, and the problems shown by these measures are discussed. Chapter 6 presents statisti-

cal methods that handle small size data. Conclution and disccusion is in chapter 7.

Chapter 2

Evaluating Chance of Geologic Success

2.1 Geologics Success

Chance of detectable hydrocarbon of a prospect is one of the basic inputs for further decision making of prospects. After drilling, the result would be either a failure or a success, which are complementary to each other. The conventional success definition refers to that the well was completed and did produce some hydrocarbons . This success term includes several different "success" in prospect assessment: economic, commercial, completion and geologic corresponding to different level of hydrocarbon volumes. Geologic success, defined by Rose (1992), is: "a well that encounters mobile hydrocarbons". Another version by Rose (2001) is: "a reservoired accumulation was found that was at least large enough to support a flowing test".

Zero hydrocarbon encountered means failure and geologic success are a concept normally complementary to the failure. However, geologic success normally involves some volume of hydrocarbon. Rose (1992) mentioned that investors are more interested in whether the well will contain enough petroleum that would cover the cost of completion of the well, rather than the presence of hydrocarbon from the geologic view. And the conventional reporting standard for exploratory success in most petroleum-producing nations is whether the exploratory well was completed for production, which can become unequivocal record (Rose, 2001). So the success data of discovering enough petroleum to complete the well would be more accessible and not depend on various economic requirements.

In order to have consistent record for pre-drill chance of geologic success and post-drill ge-

ologic success observation, introducing requirement of minimum volumetric into the definition of geologic success would be necessary. In case of *onshore* mature petroleum province, a geologic chance system should be consistent with the chance of finding enough petroleum to complete the well (Rose, 1992).

For *offshore* case, the cost of completing a well is much higher. The Geologic success is often defined as the discovery of movable hydrocarbons, which is also called technical discovery (CCOP, 2000). One example is in *Evaluation of Norwegian wildcat wells* (Ofstad et al., 2000), post drill technical discovery data was compared with pre-drill discovery probability, where technical discovery was defined as discovery larger than 2 mill Sm3 o.e. and pre-drill discovery probability was evaluated the same way as to the chance of geologic success. So the technical discovery is used as equivalent as geologic success here.

In short, generally no universal hydrocarbon volume is associated with the geologic success, as long as consistency is kept between defining geologic success and estimating probability of geologic success. The probability of geologic success is obtained near the end of the exploration process.

Petroleum exploration generally proceeds in a sequential manner (Harbaugh et al., 1995), from **sedimentary basins, petroleum systems, plays to prospects**, which can be regarded as different levels of hydrocarbon investigations suggested by Magoon and Dow (1994): investigation of sedimentary basins describe the stratigraphic sequence and structural style of sedimentary rocks; petroleum system studies describe the genetic relationship between a particular pod of generating source rock and the resulting oil and gas accumulations; investigation of plays describe the present-day geological similarity of a series of present-day traps; and prospect investigates the individual present-day trap.

In Milkov's 2015 paper, **segment** is the smallest assessment unit; segments can include individual reservoir units and compartments. As prospect represents potential petroleum accumulations. Prospects can be "combinations of several segments that occur within one common structure" (Milkov, 2015). So when the prospect has only one segment ,then the geologic success for the prospect is the same as the for the segment. Milkov (2015) mentioned that chance of geologic success for multi-segment prospects is aggregated from the estimates for segments using stochastic calculators (GeoX, REP etc.) taking the dependencies between risk factors for

different segments into account. So a failure for a segment does not necessarily lead to a failure for the multi-segment prospects containing the segment.

2.2 Obtaining Chance of Geologic Success

Common method

The most common way of estimating the probability of geologic success (P_g) is by multiplying the probability of the essential *geologic factor* of a prospect/segment. The geologic factors can be also called risk factors - independent factors that could cause the segment to fail. The prospect/segment fails so long as one of the risk factor fails. In order for the subsurface hydrocarbon accumulation to exist, essential geologic factors must coincide. According Norwegian petroleum directorate (NPD (2010)), four factors must coincide for the petroleum to be formed and accumulated:

1. A **reservoir rock** where the petroleum can accumulate
2. A **trap** so that petroleum is retained in a reservoir
3. A **source rock** containing organic material, which can convert to petroleum at sufficient temperature and pressure
4. A **migration route** that allows the petroleum to move from source to reservoir rock.

The expression for the chance of geological success is then:

$$P_g = P_1 * P_2 * P_3 * P_4.$$

However, there is disagreement between the numbers of essential geologic factors determining the P_g. 4 to 7 essential geologic factors are most commonly seen. The number of independent geologic factors can influence the probability of the succeess. As multiplication of more factors can lead to lower probability. Those geologic factors should be independent of each other. If dependency exists, the influence needs to be evaluated. Milkov (2015) summarized different

geologic factors used by different authors and how the probability of geologic success is obtained from probability of geologic factors in a table**??** presented in appendix. some alternaive methods are also listed the table. One distinct method involving historical success rate which is also in the tabel, is explained in the following section.

Alternative method

The historical success rate (number of discoveries /total number of prospects) has been directly used as a substitute or in combination of geologic chance factors in some cases. The following method suggested by Snow et al. (1996) is another way of estimating chance of success by combining success rate and estimation of geologic factors.This method of estimating the probability of finding any testable hydrocarbons before a well is drilled is illustrated in the figure 2.1. The probability is the product of the *historical success rate* or plan chance and four geologic *comparison coefficients*. As long as the hydrocarbon could be tested or flowed to surface, or the amount is larger than zero, the testable hydrocarbon is claimed.

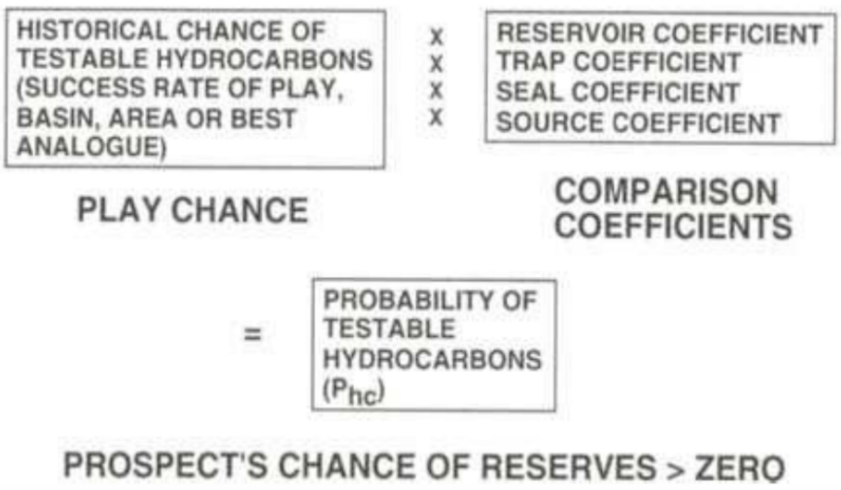


Figure 2.1: Method of determining chance of success by combining historical success rate and comparison coefficients from geologic chance factors (Snow et al., 1996)

The *historical success rate* is derived from the play by comparing the prospect to the play component by component. The success rate will be updated as the play matures. When historical data for the play is absent, success rate for the basin or the nearest analogue can be employed.

Each of the prospect's characteristic: reservoir, source, seal, and trap are compared with prospects from the historical play. The results of the comparison to the historical play are categorized into "better", "same", and "worse". Then the quality of information used for comparison is also evaluated and categorized into "direct data-high certainty", "intermediate date - moderate certainty" and "indirect data - low certainty". Then, numerical values to the comparison coefficients are assigned after considering both comparison results.

Method of using historical success rate has drawbacks. Rose (2001) pointed out that observed success rate as a proxy are a poor substitute for prospect-specific chance of success, because the characteristics of each individual prospect are unique; and the quality of the data may not be consistent with ones current estimation, and the success rate changes as the field matures.

2.3 Probability of Geologic Factor

Several elements determine a geologic factor. Probability of each geologic factor may also be determined by multiplying relevant probability of *subfactors*. But too many subfactors would lead to hopeless small chance of success, because of multiplication. In *CCOP's guidelines for risk assessment of petroleum prospects* (CCOP, 2000), each factor is determined by 2 subfactors, for example, presence of reservoir facies and effective pore volume determine the reservoir factor.

Probabilities of the geologic factor or subfactor are produced by the exploration team. The most common way is that an exploration team assign probability to each individual geologic factor after discussion. Then this probability may be reviewed or modified by another technical review team or management. Usually one explorer announces the probability first; others will modify it later (Milkov, 2015).

Those assigning are largely based on the explorers' subjective opinions. Some tools are helpful for explorers to assign probabilities by providing probability scale as reference. A "chance adequacy matrix" suggested by Rose (2001) in figure 2.2 shows probability scale corresponding to 4 dimensions: risk, confidence level, (less likely to more likely), data quality and conclusion from data. A more detailed qualitative description for the relative probability scale by textit CCOPs guidelines (CCOP, 2000) including dimensions: certainty level analogue model and geologic

model are established shown in figure in appendix.

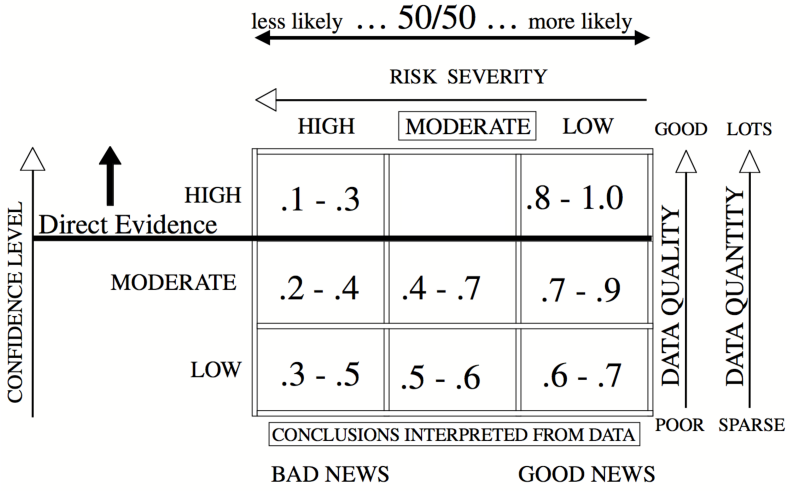


Figure 2.2: Chance adequacy matrix (Rose, 2001)

Milkov (2015) suggested that *risk table* is a more effective solution for introducing objectivity and consistency in the assigning process. A risk table shows numerical probabilities corresponding to detailed different criteria sets for individual geological factor or subfactors. Milkov (2015) has also argued that risk table is an advantageous method, making the P_g evaluation more transparent and audible, and presented such risk tables for different goleogic factors including: presence of structure, reservoir facies, reservoir deliverability, seal, mature source rocks, and migration. Such tables are also suggested by CCOP (2000): two to three dimensions of criteria are applied to evaluate the subfactors, often including data reliability and geologic properties, then numerical probabilities corresponding to different criteria sets are established in the table. One example of risk table for one geologic factor is shown in apendix.

Chapter 3

Subjectivity, Inconsistency, Eliciation and Verification

3.1 Subjective Probability

The process of estimating each geologic factor is largely subjective, although tools such as risk table improve the consistency. Subjectivity can not be avoided in the evaluation process. Since our knowledge about the subsurface world is incomplete and imperfect. The probability is best tool available which could properly reflect our lack of knowledge or information about a certain geologic parameter and provide a quantitative description of the likely occurrence of an event.

Considering probability as either the relative frequency with which a specific outcome is observed within a larger number of similar circumstances (that is, the past history of an event), or the strength of belief that some outcome will occur in the future (Harbaugh et al., 1995), is practical to understand it. People commonly believe that subjective probability assessment without precise computation is inaccurate and inconsistent, and thus should be avoided, but rather prefer a more objective procedure, involving using data in form of frequency distribution, to be employed. However, With incomplete knowledge, experts must rely on their subjective assessment of prior information, subjective beliefs are necessary for the subsurface world full of uncertainty.

In exploration assessment literatures, the subjective probability refers to personal understanding of outcome of an event; and the objective probability refers to an observed relative

frequency. According to Bratvold and Begg (2010), “a probability reflects a person’s knowledge about the outcomes of an uncertain event. Probability is a state of mind, not a state of things". So probability could only be assigned by person and thus probability are subjective. Explorations can subjectively provide probability predictions to geologic success or geologic factors by considering information acquired by geotechnical investigations, past experience and knowledge, or results from statistical calculations based on empirical data, or a combinations of them. The probability can describe rational expectations of a future event.

3.2 Inconsistency in Assigning Probability

Extensive psychological research has shown that people, even experts, tend to find it difficult to assess probabilities; to simplify this task they use heuristics, most often leading to poorly calibrated and biased assessments (Kahneman and Tversky, 1982). Especially our knowledge and data are limited compared to the complexity of the subsurface world. It is difficult facing complex problems with multiple determine factors. Experts are susceptible to bias and heuristics (rules of thumb or mental shortcuts), both on an individual basis and group basis.

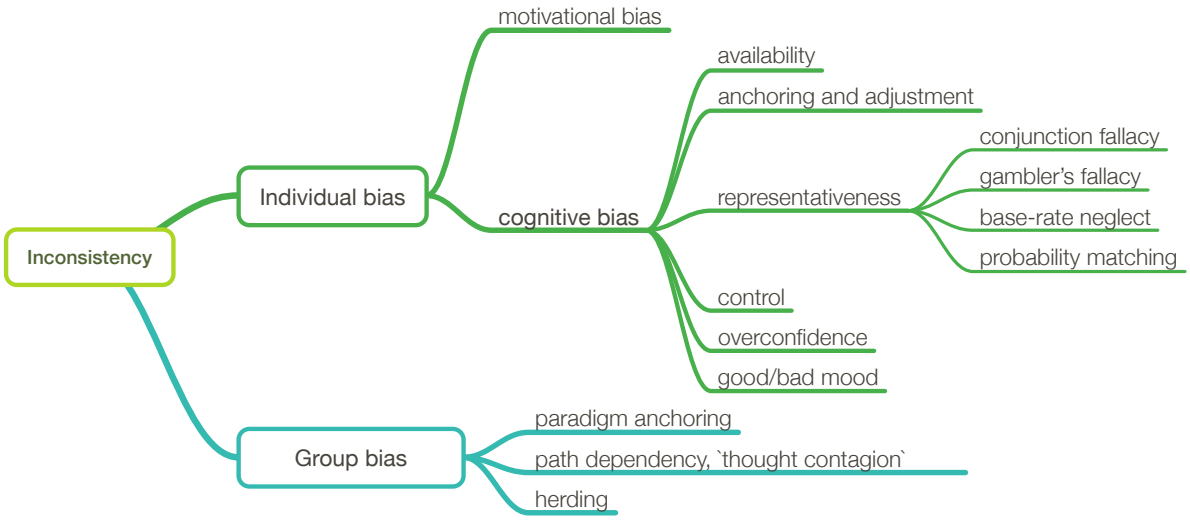


Figure 3.1: Typical bias and errors

Individual bias can be categorized into motivational bias and cognitive bias. Motivational bias reflects the interests and environments of explorers, including, such as, career/performance

targets pressure, need to influence decisions and etc. Cognitive biases are caused by using heuristics. Several biases are particularly common in oil and gas industry, such as, availability, anchoring, overconfidence and etc (Bratvold and Begg, 2010). Fig. 3.1 summarizes typical human bias and errors presented by Baddeley et al. (2004) who explore cognitive issues surrounding prior information baseed on probailistic judgment in their paper. As discussed in previous chapter, explorers produce the probability in a group. The group interactions generate and perpetuate more complex form of bias. Three approaches to explain group bias are also listed in fig. 3.1.

Rose (2001) has summarized characteristics patterns of predictive bias in estimating chance of success. For those high risk new field wildcat (NFW) prospects, companies and explorers tend to be seriously overoptimistic in predicting chance of success: for NFWs having a predicted chance of success of 10% or less, less than 1% resulted in discoveries. For intermediate-risk (20-35% range) ventures, actual success rates were generally matched predictions. For low-risk (35-60% range) ventures, more of those ventures were successful, so the predictions were conservative. For high-confidence (60-90%) ventures, actual results were notably lower. The predicted probabilities were biased. One of the reasons is that they are tempered by subjective appraisals of reviewers and managers.

3.3 Elicitation

Some of those bias can be handled and partly reduced by elicitation method. Elicitation is the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a (joint)probability distribution for those quantities (Garthwaite et al., 2005). There are several probabilitiy elicitation protocols: the Stanford Research Institute (SRI) assessment has been the most influential in shaping structured probability elicitation. The expert is engaged in a five-stage process including:

> **motivating** the experts with the aims of the elicitation process; **structuring** the uncertain quantities in an unambiguous way; **conditioning** the expert's judgement to avoid cognitive biases; **encoding** the probability distributions; and **verifying** the consistency of the elicited distributions.

Curtis and Wood (2004)

the main aim of the SRI protolcal is to help explorers to avoid psychological biases(Hora, 2007).

SRI protocol was designed for single experts, Hora (2007)mentioned Sandia protocol which was deigned to bring multiple experts. The sandia protocol consists of two meetings. After some period individual study period, first meeting is hold including presentation of the issues, discussion by the experts ,and a training session and feedback. Second meeting includes discussion about methods and data sources used and individual elicitation.

More or less elicitation (MOLE) proposed by Welsh et al. (2008) shows a benefit in both the precision and the accuracy of elicited ranges. MOLE is a heuristic-based elicitation method encourages people to consider more values by asking them to make repeated relative judgments.

Many literatures handle elictaitoin methods to tekle and reduce bias, though few elictations have been used in the geoscience, and Curtis and Wood (2004) has presented pertinent elicitation mtheods for geoscience. Elicitation theory mitigates the effects of bias, but a method to estimate reliable uncertainties on expert judgements remains elusive (Baddeley et al., 2004). Forecast verification provides multi-facets quality evaluation of probability prognosis, which may improve the calibration of probability.

3.4 Forecast Verification

An elicitation is done well if the probability distribution represents the expert's knowledge, regardless of how good that knowledge is, which corresponds to *normative goodness*. Two standards of goodness of the assessors are identified by Winkler and Murphy (1968): *Normative goodness* and *Substantive goodness*. Normative goodness concerns expertise in probability assessment and requires probabilities correspond to experts' judgment. Substantive goodness concerns expertise in the geotechnical or geology domain in which assessments are made and requires probabilities correspond to reality.Thus, Goodness of the probabilities produced by explorers comes from firstly the quality of explorer's exploration related knowledge, and secondly the accuracy with which that knowledge is transformed into probability. Measuring the goodness would encourage explorations to make "honest" judgment, provide means to evaluate explorers' prediction performance, and improve their predictions in the end. Varying statistical

measures can determine the goodness of the assessor or the assessment.

Though there are lots of inconsistencies in geologic subjective evaluations, there is potential for improvement of those probability predictions, as appropriate procedures and framework accompanies the evaluations. Meteorologist, for example, make comparatively accurate predictions, one reason is that forecast verification is extensively developed in meteorology, so that their forecast is systematically studied along with actual observations.

> A *verification measure* is any function of the forecasts, the observations, or their relationship, and includes for example the probability of the event being observed (the base rate), even though this is not concerned with the correspondence between forecasts and observations.
>
> Jolliffe and Stephenson (2012)

Forecast verification measures concern not only the relationship between observations and forecasts, but also the forecast itself or observation itself being subject of the study. There is extensive number of verification measures.

Post-drill analysis is increasingly used to improve the quality of predictions and estimations made by explorationists, however statistical methods being used are very limited. Verification can be part of elicitation process that helps to make calibrate probability and has many other benefits. Forecast verification has been developed quite extensively in weather and climate forecast area and is also popularly used in medical diagnostic tests and economic forecasts. The benefits identified by Jolliffe and Stephenson (2012) using verification measures in different disciplines would be also great incentives for exploration ventures: A numerical measure of how well are those geologic assessments would be very helpful from a administrative point of view, which can be used for management or strategy development purpose; a greater understanding of the problems under assessing; possibility of improved understanding of the underlying geological or geophysical processes by use of more detailed verification measures. A detailed understanding of strengths and weakness of their forecasts; and more concrete information about the quality of the forecasts supporting rational decision makings.

Chapter 4

Scalar Verification Measures

Measures in this chapter give scalar value or score for the forecast's quality. These measures are also fundamental concepts that describe attributes of forecast verification. Except common forecast verification measures, most of the following measures and formulations are presented by Bradley et al. (2003a) in their *Distributions-oriented verification of probability forecasts for small data samples*. These measures are applied to the first dataset, 184 pairs of chance of success and results of prospects on NCS from NPD in the end. The results of scaler measures may be hard to understand at first, reader may read graphical measures in chapter 5 firstly, and come back to chapter 4 for relevant forecast verification attributes.

4.1 Summary Measures

Summary measures or measure-oriented measures concern some overall quality of forecast, including common statistical indicator such as mean and mean error, and different scores

Mean Error /Bias

The mean error or the unconditional bias measures the average difference between a set of forecast and corresponding observations.

$$ME(f,x) = \frac{1}{n}\sum_{i=1}^{n}(f_i - x_i) \tag{4.1}$$

n represents the sample size; This measure is simple and often seen. But the pitfall is obvious that the score does not measure the correspondence between forecasts and observations. It is possible to get a perfect score when there is actually bad correspondence between forecasts and observations. Range $-\infty$ to ∞. Perfect score: 0.

Mean Squared Error /Brier Score

MSE gives a numerical result to a single forecast or to a set of forecast. Specifically, the brier score measures the mean squared difference between the predicted probability to the possible outcomes and the actual outcomes. Thus, the lower the brier score is for a set of predictions, the better the predictions are calibrated.

$$MSE(f,x) = \frac{1}{n}\sum_{i=1}^{n}(f_i - x_i)^2 \tag{4.2}$$

Range 0 to 1. Perfect score: 0.

Skill Score(SS)

Skill score(SS) or Brier skill score measures the performance of one forecasting system relative to a reference forecast in terms of the Brier score (BS).

$$SS = \frac{BS - BS_{reference}}{0 - BS_{reference}} \tag{4.3}$$

So, a skill score of 0 indicates no skill and a skill score closer to 1 is preferred. It should be noted that this score is not stable on small data sets, and for larger the number of sample needed for more rare events (CAWCR, 2014). As the base rate prediction is commonly used as the "unskilful" reference, the equation is equivalent to:

$$SS = \frac{\sigma_x^2 - MSE(f,x)}{\sigma_x^2} \tag{4.4}$$

Range $-\infty$ to 1. Perfect score: 1.

4.2 Distributoin-Oriented (DO) Measures

In the late 1980s, Murphy and Winkler (1987) proposed a general framework of forecast verification called the Distributions-Oriented (DO) approach. The DO approach is based on the joint distribution of forecasts and observations, indicating various facets of the forecasts quality and allows the user to evaluate forecast performance on specific situations.

4.2.1 Joint Distribution and Factorization

The **Joint distribution** of forecasts and observations can be represented as $p(f,x)$, where f represents the probability forecasts and x represents the observation, and $p(f,x)$ is the joint probability of f and x. $p(f,x)$ contains information about the forecasts, the observation, and the relationship between the forecasts and observation. For verification purpose, $p(f,x)$ can be interpreted as an empirical relative frequency distribution based on a sample of past forecasts and observations. Any joint distribution can be factored into a conditional distribution and a marginal distribution in two ways.

Calibration-Refinement Factorization (CR)

$$p(f,x) = q(x|f)s(f) \tag{4.5}$$

$p(f|x)$ is the conditional distribution of observation given each forecast value (called *"calibration/reliability"*) and $p(f)$ is the marginal distribution of the forecasts (called *"sharpness/refinement"*).

Likelihood-Base Rate Factorization (LBR)

$$p(f,x) = r(f|x)t(x) \tag{4.6}$$

$p(x|f)$ is the conditional distribution of forecasts given each possible observation (called "*likelihood*") and $p(x)$ is the marginal distribution of the observations (called "*base rate*").

4.2.2 Calibration Refinement Measures

The calibration-refinement factorization conditions on forecasts. Given a forecast, certain aspects of the distribution of the Observation x can describe some quality of the forecast perfor-

mance.

Reliability or calibration (also called *Type 1 conditional bias*) measures the degree to which the forecast probabilities correspond to the conditional frequency of occurrence of the event. For example, if the event occurs on 30% of the occasions when 30% has been the forecast, this forecast is said to be reliable. Reliability can be described as the bias of the observation given a forecast f:

$$REL = E_f(\mu_{x|f} - f)^2 \tag{4.7}$$

where E_f is the expected value with respect to the distribution for the forecasts and $\mu_{x|f}$ is the expected value of observations conditioned on the forecasts. So the smaller the relabiity value, the better quality of the forecast is made.

Resolution: The ability of the forecast to distinguish situations with distinctly different frequencies of occurrence, E.g., the observed frequency of occurrences when predictions are 60% is compared with observed frequency of occurrences when predictions are 30%, there should be difference when there is resolution. Resolution describes the degree to which the mean observation for a specific forecast f is different from the unconditional mean of observatoin (base rate):

$$RES = E_f(\mu_{x|f} - \mu_x)^2 \tag{4.8}$$

Forecasts with larger differences,i.e., higher resolution are more desirable. Even if the forecasts are wrong, the forecast system has resolution if it can successfully separate one type of outcome from another.

CR Decomposition of MSE: MSE can be decomposed into components containing both reliability and resolution. The so-called calibration refinement decomposition:

$$MSE_{CR}(f,x) = \sigma_x^2 + REL - RES \tag{4.9}$$

$\sigma_x^2$is the variance of observation, measuring the inherent **uncertainty** of the observation.

Relative measures: substituting the CR decomposition of MSE into the skill score:

$$SS = \frac{RES}{\sigma_x^2} - \frac{REL}{\sigma_x^2} \tag{4.10}$$

Then, dividing the RES and REL terms with variance of observation gives us the relative resolution and relative reliability. Then the two terms on left side of equaton 4.10 are normalised reliability and resolution against the uncertainty of event .

4.2.3 Likelihood-Base Rate Measures

The likelihood-base rate (LBR) factorization conditions on the observation. Given a specific observation (occurrence $x = 1$ or non-occurrence $x = 0$), certain aspects of the distribution of the forecast f can describe some quality of the forecast performance.

Discrimination: a measure of how well the forecasts discriminate between events or non-events, or ability of the forecast to discriminate among observations, that is, to have a higher prediction frequency for an outcome whenever that outcome occurs. Discrimination is evaluated by measuring the difference between the two conditional distributions of forecast probabilities, $p(f|x = 1)$ and $p(f|x = 0)$ for binary event.

$$DIS = E_x(\mu_{f|x} - x)^2 \tag{4.11}$$

Discrimination is evaluated by measuring the difference between the two conditional distributions of forecast probabilities, $p(f|x = 1)$ and $p(f|x = 0)$ for binary event. For dichotomous event:

$$DIS = (1 - \mu_x)(\mu_{f|x} - \mu_f)^2 + \mu_x(\mu_{f|x=1} - \mu_f)^2 \tag{4.12}$$

Forecasts with large differences, i.e., high discrimination are more desirable.

Type 2 conditional bias describes the bias of the forecast given the observation. One measure of this bias is:

$$B_2 = E_x(\mu_{f|x} - x)^2 \tag{4.13}$$

For dichotomous event:

$$B_2 = (1 - \mu_x)\mu^2_{f|x=0} + \mu_F)^2 + \mu_x(\mu_{f|x=1} - 1)^2 \tag{4.14}$$

LB Decomposition of MSE: MSE can be decomposed conditioning on the observation. The

so-called likelihood - base rate decomposition:

$$MSE_{LBR}(f,x) = \sigma_f^2 + B_2 - DIS \tag{4.15}$$

Relative measures: substituting the LBG decomposition of MSE into the skill score:

$$SS = \frac{RES}{\sigma_x^2} - \frac{REL}{\sigma_x^2} \tag{4.16}$$

Then, dividing the DIS and B_2 terms with variance of observation gives us the relative Discrimination and relative B_2.

4.3 Estimation of Measures- Continuous Approach

Measures as mean, variance can be directly estimated from the sample datasets. The conditional mean $\mu_{f|x=1}$ and $\mu_{f|x=0}$ in Discrimination and Type 2 conditional bias can be estimated by equation 4.17 and 4.18. Let the sample dataset be partitioned into two sets, let $f_j^0, j = 1,\ldots,N_0$ be the cases when observation $x = 1$ and let $f_k^1, k = 1,\ldots,N_k$ be the cases when observation $x = 0$:

$$\mu_{f|x=0} = \frac{1}{N_0}\sum_{j}^{N_0} f_j^0 \tag{4.17}$$

$$\mu_{f|x=1} = \frac{1}{N_1}\sum_{k}^{N_1} f_k^1 \tag{4.18}$$

The $\mu_{x|f}$ term in Reliability 4.7 and Resolution 4.8 can not be directly estimated from the sample datasets. However for large datasets, it is okay to use approximate value of $\mu_{x|f}$ calculated from *Contingency table method.* For the contingency table method: firstly prognosis data are binned into several categories; then a contingency table can be built to acquire joint distribution , marginal distribution and conditional distributions of observations and prognosis. The distribution of mean observation given a certain forecast category $p(x = 1|f)$ are obtained from contingency table. The $P(x = 1|f)$ can be used as a approximation value for the $\mu_{x|f}$, Then corresponding value for Reliably and and resolution can be obtained. For small datasets, contingency table method my distort or hide the original information of prediction quality. Statistical

methods can be applied to handle small dataset which are specifically discussed and presented in chapter 6.

Table 4.1 summarizes the definitions of attribute (scalar measures) of verification and corresponding forecast or observation distributions and relevant masseuses.

4.4 Verifcation for Prognoses of 184 Prospects on NCS (1998-2007)

184 Prospects on NCS	
μ_f	0.2994
μ_f	0.4022
ME	-0.1028
MSE	0.2238
X_variance	0.2417
f_variance	0.0314
Skill score	0.0743
Discrimanation	0.0035
Type 2 bias	0.1961
Reliability	0.0108
Resolution	0.0275

Table 4.1: Summary of meausres for the entire 184 pairs of probabilities of discovery and resutl (1998-2007) on NCS

There are 184 pairs of prognoses and results of prospects drilled between 1998 and 2007 on the Norwegian Continental Shelf in the first dataset presented in chapter 1. Among three datasets, only this first one provides complete original predictions and corresponding data, so that all the scalar measures can be directly applied. Prognosis are probabilities, and results are either discovery ($x = 1$) or failure($x = 0$).

The summary scores of the entire 184 pairs of data are shown in table 4.1.

The scores and measures for the entire data range give an general overview of the probability of discovery evaluation performance, such as the mean obersvation is 10% higher than the mean pre-drill prediction. However, it would be very dangerous to make judgements about

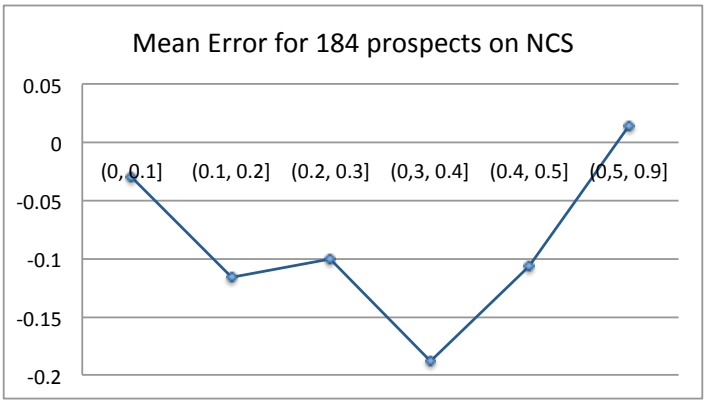
Attribute	Definition	Basic distribution(s)	Graphs and measures
Sharpness (refinement)	Degree to which probability forecasts approach zero and 1; "spread" of distribution of forecasts	$p(f)$	• Histogram of $p(f)$ • Variance of forecasts, σ_f^2
Resolution	Difference between $\mu_{x\|f}$ and μ_x, considered over all values of f	$p(x\|f)$, $p(x)$	• Resolution component of Brier Score • Attributes diagram
Discrimination	Degree to which forecasts discriminate between occasions when x=1 and occasions when x=0	$p(f\|x)$	• Discrimination diagram (plot of likelihood functions) • Difference in conditional means: $\mu_{f\|x=1} - \mu_{f\|x=0}$
Bias	Difference between mean forecast and mean observation	$p(f)$, $p(x)$	Mean Error (ME): $ME = \mu_f - \mu_x$
Reliability (Calibration)	Degree of correspondence between conditional relative frequencies, $p(x\|f)$ and f, considered for all values of f	$p(x\|f)$	• Reliability diagram • Attributes diagram • Reliability measure from Brier score decomposition
Accuracy	Average degree of correspondence between f and x	$p(f,x)$	• Brier score = MSE • Other scores
Skill	Accuracy of forecasts relative to accuracy of forecasts based on a standard of comparison (e.g., climatology)	$p(f,x)$	• Brier skill score, BSS • Correlation, $\rho_{f,x}$, measures potential skill • ROC Area

Figure 4.1: Attributes of forecast quality for probabilistic forecasts (Gofa(HNMS), 2010)

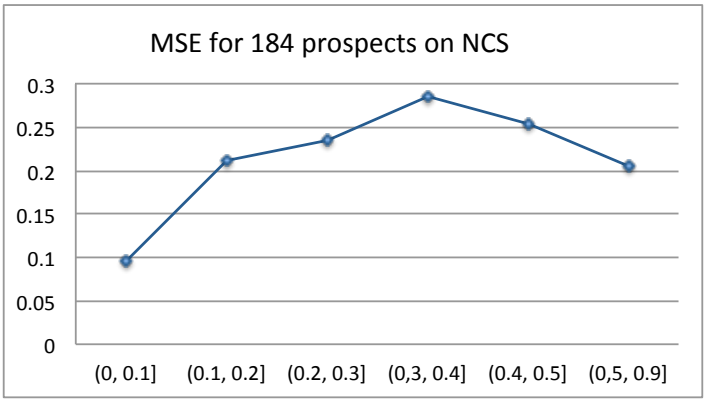
the prognosis performance. After binning the prognosis data into several categories, more detailed information of prognosis performance can be obtained by applying verification measures.

Thus, the prediction probabilities are binned into 6 categories (0.10%] (10%, 20%], (20%, 30%], (30%, 40%], (40%, 50%], and (50%,90%]. Since there are not many prognoses larger than 50%, all prognosis larger than 50% are put into one bin, the following section shows results of scalar measures for each prediction category.

Absolute Scalar Measures



(a) Mean error

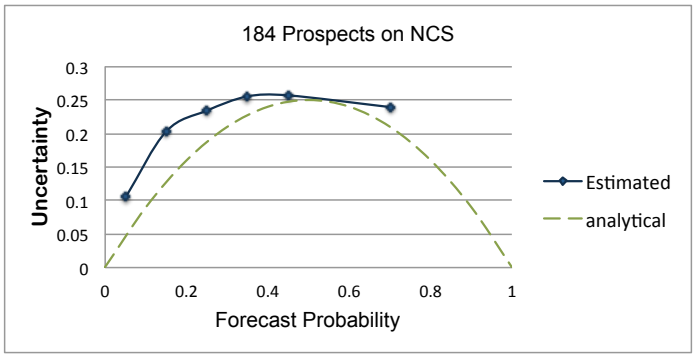


(b) Mean Squared Error

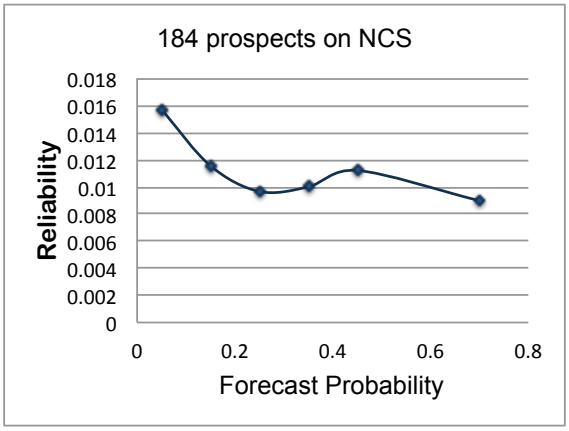
Figure 4.2: Mean error and MSE for 184 prospects on NCS

Mean error measures the bias and MSE measure the accuracy. As shown fig. 4.2a, ME -the differences between predicted probability and the actual discovery rate are negative for five bins and positive for one bin, and has greatest negative value for the bin (30%,40%). So the prognoses are generally pessimistic and are mostly conservative around 35%. MSE shows the accuracy is best for predicted probability less than 10%, and worst around 35% in fig. 4.2b.

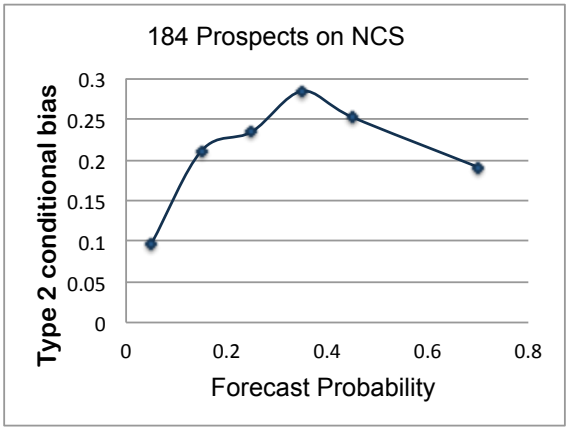
The *uncertainty*,i.e., the variance of the observation indicates the inherent variance of the



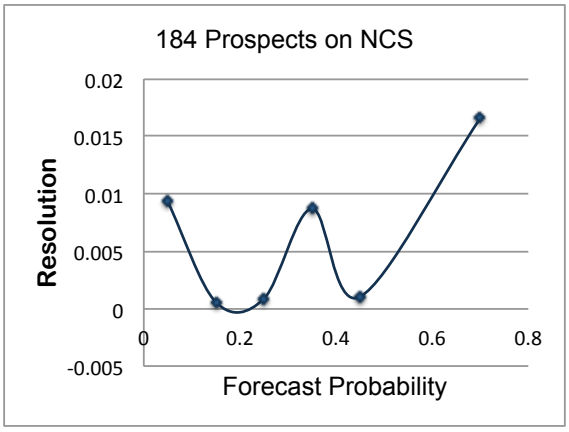
(a) Estimated uncertainty and analytical uncertainty



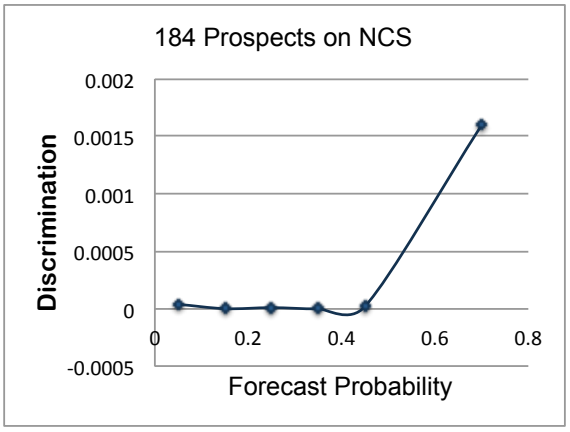
(b) Reliability



(c) Type 2 conditional bias



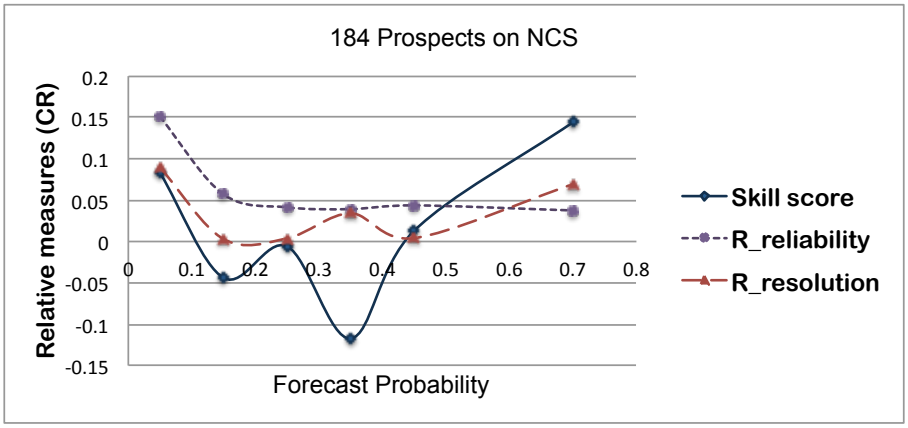
(d) Resolution



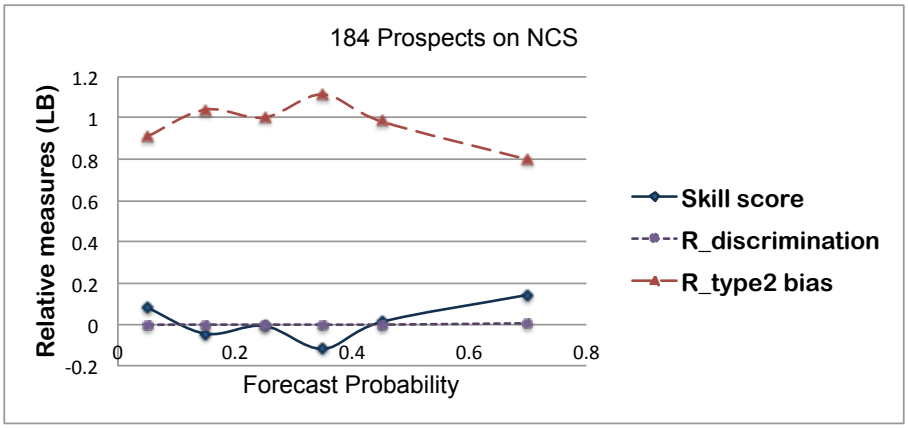
(e) Discirimation

Figure 4.3: Absolute scalar measures for 184 pairs of probabilities of discovery and results

problem are shown in fig. 4.3a. The analytical uncertainty comes from the formula $\mu_x(1-\mu_x)$, where$p$ indicates the mean observation,that is, base rate. The value of the uncertainty ranges from 0 to 0.25. If the mean observation is 0.5, there is more uncertainty inherent in the forecasting situation and uncertainty will be closer to 0.25.



(a) Relative calibration-refinement measures and skill score



(b) Relative likelihood-base rate measures and skill score

Figure 4.4: Relative scalar measures normalised by uncertainty and skill score for 184 prospects on NCS

The absolutes measures depend on mean observation μ_x of each forecast bin. So when evaluating the probability performance, relative meaesures should be considered together with absolute measures normalized by the uncertainty term. Smaller *reliability* values are preferred, so it is mostly unreliable for probabilities less than 10% and also unreliable in category(40%, 50%) among the 6 categories, shown in 4.3b and fig. 4.4a. *Resolution* are worst on the two ends and the predictions are best resolved for predictions around (30%,40%], shown in fig. 4.3d and fig. 4.4a , as larger resolution value is desired. Larger *dsicrimination* indicates better ability of the prediction to discriminate between dry and success, so better discrimination ability for predictions over 50%, shown in 4.3e and fig. 4.4b, but this high resolution may be caused by the wide range of predictions values included in this forecast bin ,which remains to be investigated. Type 2 conditional bias is not normally explained in forecast verification, which will not be discussed

more here as well. In fig. 4.4, The *skill score* shows that the forecast has lowest score around 35% ,and has most skill for probabilities over 50%.

These scaler measures provides exact scaler values ,but may not be easy to understand. The graphical measures in the next chapter can provide more intuitive and direct information. Forecast verification terms as reliability, resolution, discirimation and etc can be better understood accompanied by graphical measures.

Chapter 5

Graphical Verification Measures and ROC Analysis

Graphical diagrams based on marginal and conditional distributions of predictions and observations provide more details of the forecast performance. Sharpness histogram, reliability diagram and discrimination diagram together provides a more complete picture of the forecast quality. ROC analysis can potentially be useful for providing information for further decision-makings. All these measures are based on joint, marginal, and conditional distribution of observations and forecasts, which can be acquired by binning prediction data and using *contingency table method* mentioned in chapter 4.3. Measures in this chapter can be applied to data generated from contingency tables, so no oringinal pairs of data are required. These measures will be applied to all 3 sets of data presented in chapter 1 after explaining each measure.

5.1 Attributes Diagram & Sharpness Histogram

Sharpness (refinement): sharpness depends only on the forecast and is a characteristic, which reflects the degree of forecast definiteness. A forecaster of perfect sharpness would only give 0 and 1. If the same forecast is always given, then the forecasts are said not to be sharp (Murphy and Winkler, 1987). So unvarying forecasts have zero sharpness.

The frequency of forecasts or the number of forecasts in each probability bin of a histogram is called **sharpness histogram**. The histogram is often in the attributes diagram and shows the

sharpness of the forecast. Forecast system that is capable of predicting probabilities different from the observed frequency of the event are said to exhibit sharpness.

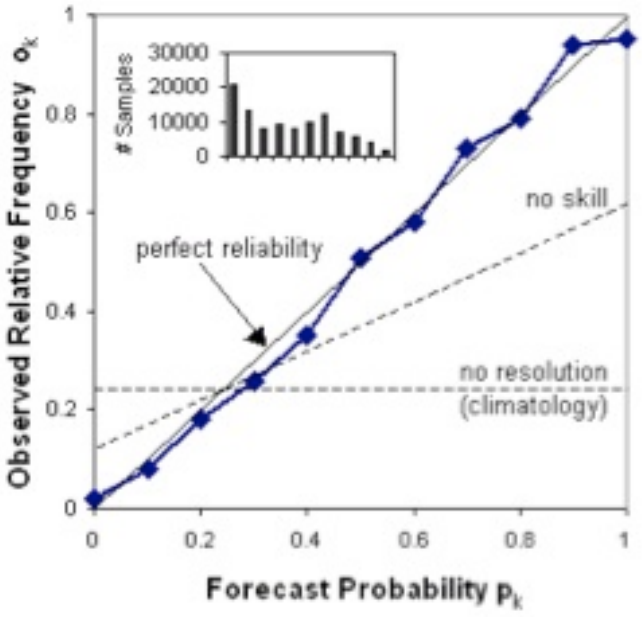
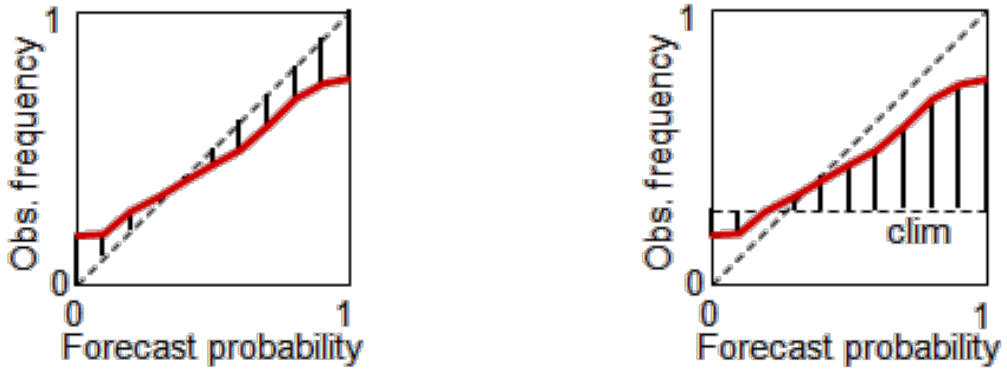


Figure 5.1: Attributes diagram and sharpness histogram (CAWCR, 2014)

Attributes diagram (sometimes called "reliability" diagram) graphically assess reliability, resolution of probability forecasts for dichotomous outcomes, and plots the observed frequency against the probability forecast. Firstly the forecast probability is divided into several bins, then the curve is determined by the average forecast of each bin on X-axis and the corresponding relative observed frequency within each forecast bin, i.e., $p(x = 1|f)$ on Y-axis. One example is shown in fig. 5.1

In a perfect reliable system the forecast probability is equal to the observed frequency, so the graph is the 45-degree diagonal line (reliability is zero). So the diagonal line is the **perfect reliability line**. If the curve below the diagonal line, the probabilities are overestimated, and if the curve above the diagonal line, the probabilities are underestimated. The horizontal line refers to **no resolution line**. A forecast of mean observation does not discriminate between observed and non-observed, and has no resolution. The **no resolution line** is constructed by plotting the base rate (mean observation). The flatter the curve, the lesser resolution it has. The **no skill line** is halfway between no resolution line and perfect reliability (diagonal) line and is where reliability and resolution are equal. Typical cases of the sharpness histogram and

attributes diagram are presented and explained in appendix.!!!!Which helps one to understand.



(a) Reliability term

(b) Resolution term

Figure 5.2: Reliability and resolution in attributes diagram (CAWCR, 2014)

The *realibility* mesaures the difference between the forecast and the mean obersvation associated with that forecast value, over all of the forecasts. Graphically, it measures the mean square distance of the curve in the attributes diagram to the diagonal line as shown in fig. 5.2a. The more reliable the forecast system, the reliability value or the shaded area is closer to 0.

The *resolution* term measures the mean square distance of the graph line to the no resolution line shown in fig. 5.2. The resolution term is large if there is enough resolution to produce very high and very low probability forecasts. Resolution is independent of reliability and is only a measure of how the different forecasts are classified or resolved by a forecast system.

Applied to 3 sets of data

Predicions of all 3 sets of data are categorized into 6 bins. Attributes diagram and sharpness histogram are built from contingency table made from theses 3 sets of data.

As u-shaped distribution of forecasts is deemed as good sharpness, so all three sets have poor sharpness. Also by reading the sharpness histogram, one learns distribution of the number of prospects being predicted in each probability category. As shwon in fig. 5.3, 5.4, and 5.5, probabilities larger than 50% take very small proportion of the total forecasts: with around 6% for BP's 805 targets, 11% for NPD 184 prospects on NCS, and 17% for NPD 118 prospects on NCS which is a bit more than the other two. The number of 'risky prospects', i.e., probabilities smaller

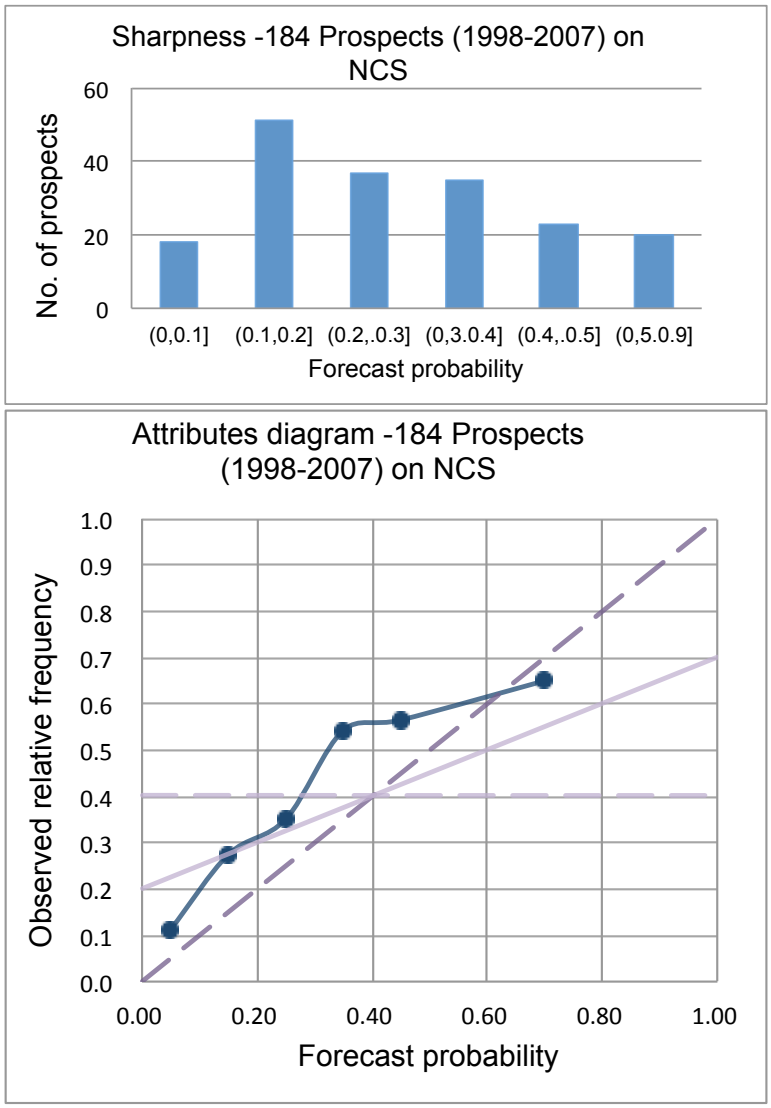


Figure 5.3: Attributes Diagram & Sharpness Histogram for 184 prospects (1998-2007) on NCS

10%, takes around 10%, 10% and 24% for the 184 on NCS, 118 on NCS, and BP 805 respectively. The number of risky prospects for BPs 805 are significantly more than the other two.

A lot of information can be read from attributes diagram. For 184 wells (1998-2007) on NCS in fig. 5.3, Probabilities are generally underforecasts, except in the last categorry (predctions larger than 50%). For 118 prospects (1990-1997) on NCS in fig. 5.4, for probabilities between 15% and 35%, a bit underforecast is observed or in another word, the forecast is a bit pessimistic; and for forecasts larger than 40%, apparent overforecast or overconfidence is observed. For forecasts (40%, 50%) and forecast lager than 50%, the forecast are almost on the no skill line, indicating not much skills in the forecasts. It can be summarized that probabilities below the base rate

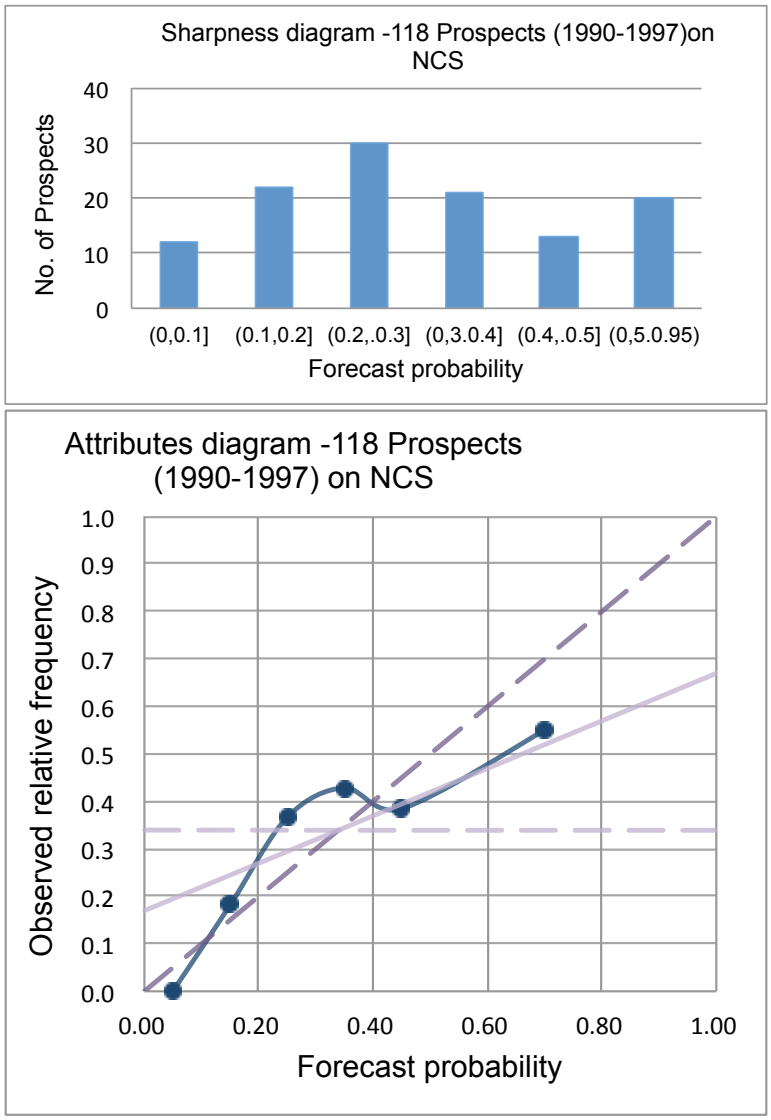


Figure 5.4: Attributes Diagram & Sharpness Histogram for 118 prospects (1990-1997) on NCS

(mean observation) are underforecast and probabilities above the mean observation are over-forecast. This kind of forecasts that tend towards to the mean observation is typcial: by only comparing the mean obsertion with mean forecast, forecast may match well with observation. However, differences between diffrernt forecast categories can not be detected by simply comparing the mean. BP'S 805 well(1983-1997) are bestly calibrated among the 3 datasets shown in fig. 5.5, as the curve are mostly close to the perfectly reliability (calibration) line. The predictions are a bit under-forecast.

For the NPD 118, the cure in fig. 5.4 alomost lies on the no skill line for probabilites larger than 40%, idicating no skills in the chance of success pronoses.

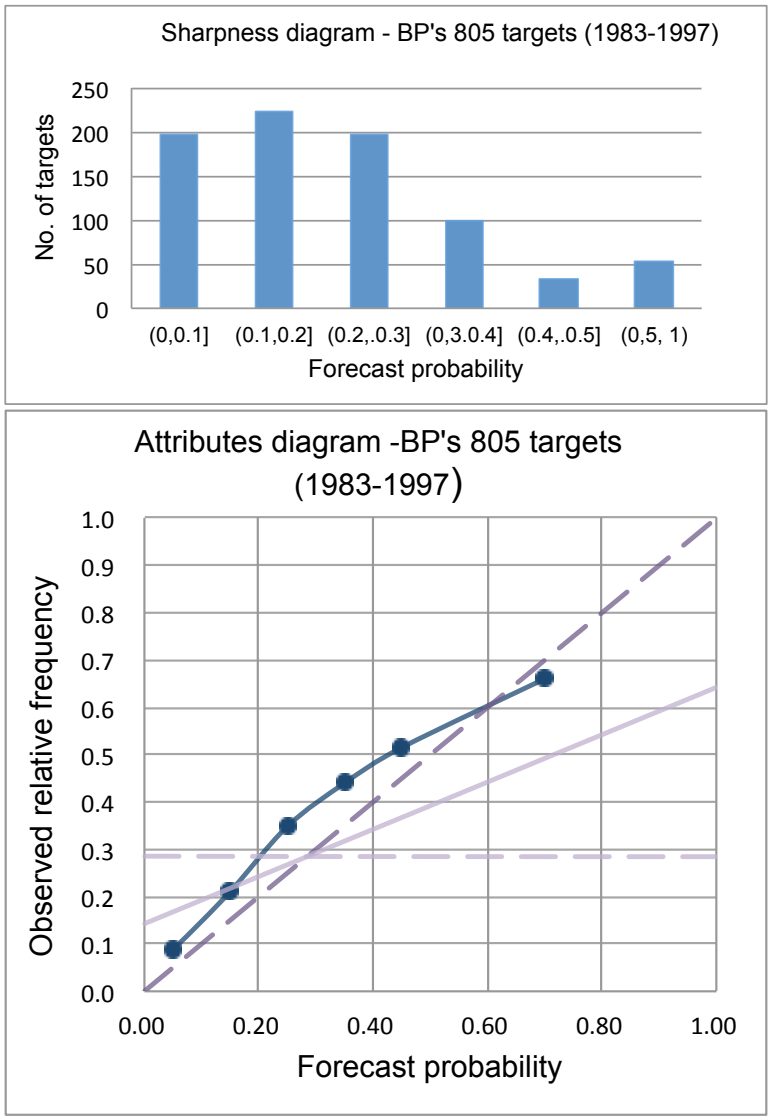
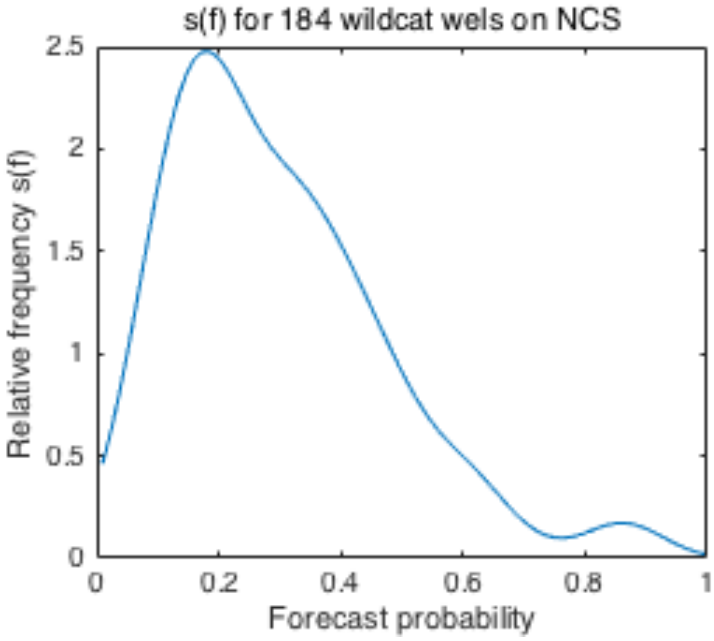


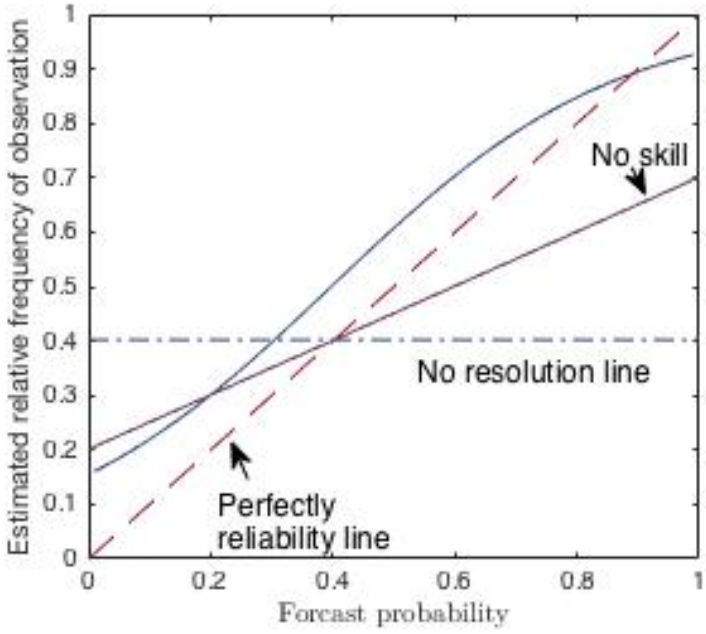
Figure 5.5: Attributes Diagram & Sharpness Histogram for BP's 805 targets (1983-1997) on NCS

For all three sets, overconfidence or over-forecast exists for forecasts larger than 50%. Combined with sharpness histogram, probabilities of BP's 805 wells are also least centered on the mean observation among the three sets while probabilities of 118 on NCS are mostly centered around the mean observation 34% which reveals potential 'anchoring' bias. For the NPD 118, the cure in fig. 5.4 almost lies on the no skill line for probabilities larger than 40%, indicating no skills in the chance of success proposes. The curve are mostly smooth For BP 805 as most data are contained.

5.2 Sharpness and attributes diagram estimated by continuous approach



(a) Marginal distribution of predictions (Sharpness)



(b) Conditional distribution $\mu_{x|f}$ (Attributes Diagram)

Figure 5.6: Attributes Diagram & Sharpness for 184 prospects (1998-2007) on NCS by continuous method

Either the measures are estimated directly from complete original pairs of data, so-called **continuous approach** described in chapter 4 section 4.3; or the **contigency table method**, also called *discrete approach*, where the data are firstly binned into contigency table ,measures are then based on distribution acquired from the contingency table.

For the 184 prospects on NCS, original pairs of data are available, thus continuous approach can be applied. To visualize one-dimensional conditional distribution of prediction, histogram of sharpness is one way; the distribution of forecast can also be estimated by the Kernel density estimation (KDE) method described in next chapter. In fig. 5.6a, the curve is relative frequency of 90 probability data points showing the sharpness of predictions. Conditional distribution $\mu_{x|f}$ obtained by logistic regression (LR) method described in next chapter, can become an attributes diagram by adding no skill line, no resolution line and perfect reliability line, in fig. 5.6. some differences can be observed compared with fig. 5.3 acquired from contingency table method.

5.3 Discrimination diagram

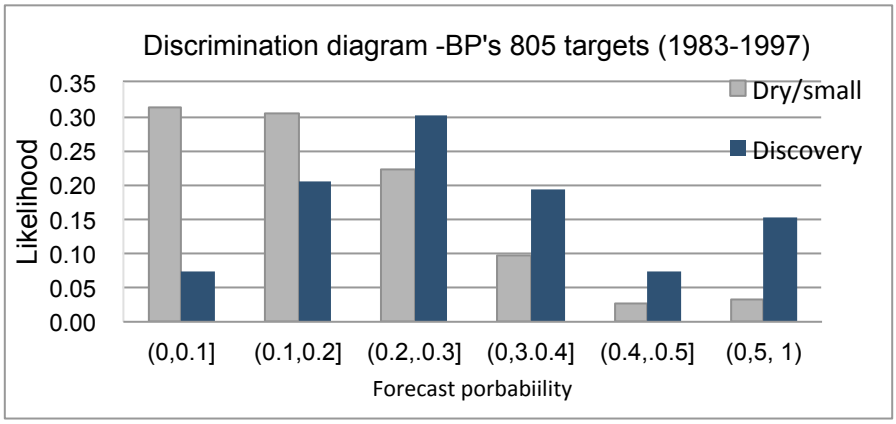
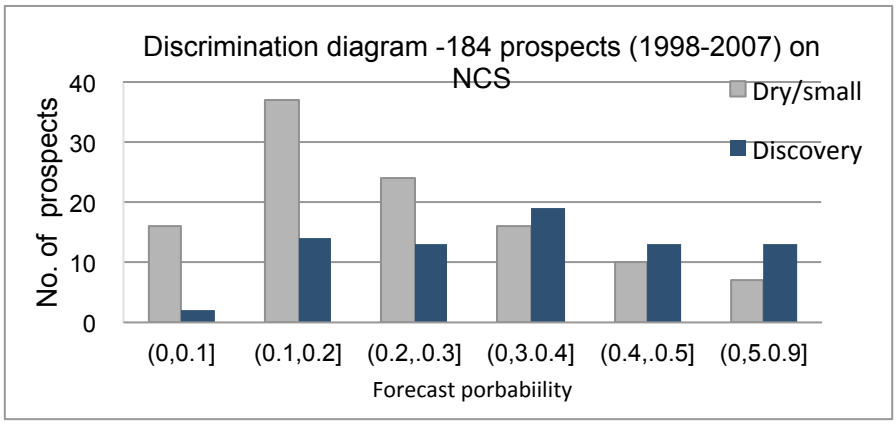
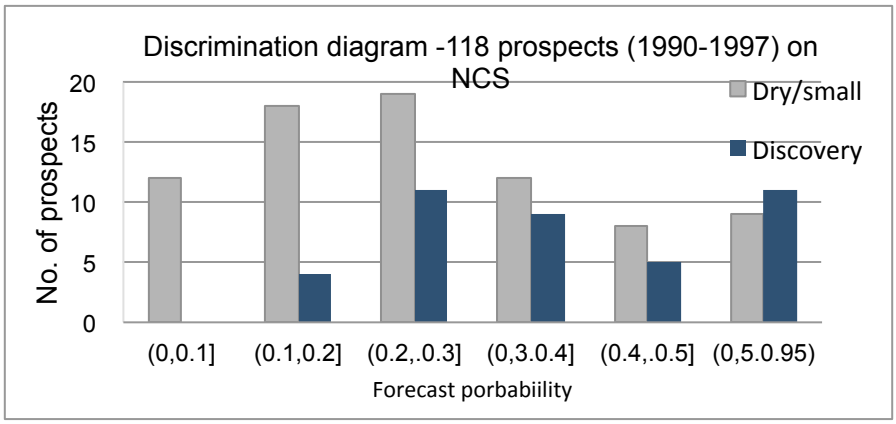


Figure 5.7: Discirimination diagram for BP's 805 targets,with likelihood ratio

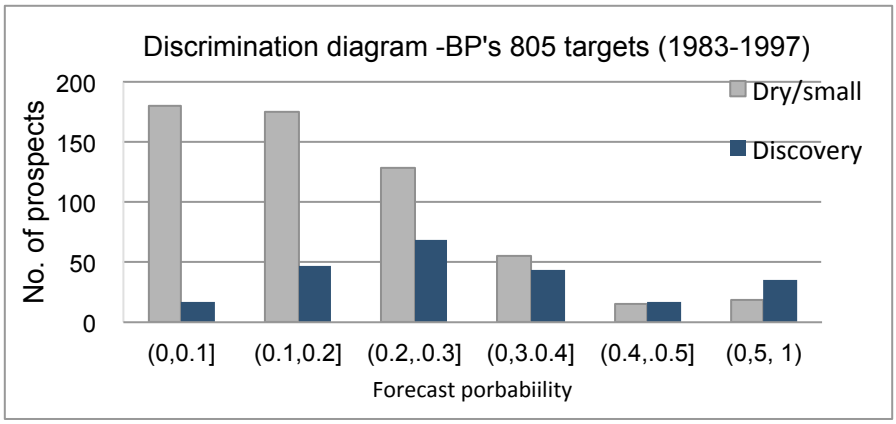
For a specified observation category, the distribution of forecast is shown in discrimination diagram. Conditional probabilities $p(f|x=1)$ and $p(f|x=0)$ are called the likelihoods associated with the forecast. Discrimination diagram is constructed by plotting the likelihood of each forecast probability when the outcome is observed and when the outcome is not observed. If $p(f|x)$ is very similar for different x, the forecast is not very discriminatory; if the likelihoods are very different for different x, the forecast is much more discriminatory (Murphy and Winkler,



(a)



(b)



(c)

Figure 5.8: Discrimination diagrams

1987). Perfect discrimination is when there is no overlap between the distributions of probabilities for observed and not observed, so it is poor discrimination when the two distributions overlap much.

When there is much data, likelihood is a good representation, for example for the BP's 805

targets in fig. 5.7. For our problem, there are mostly not much data points, so number of prospects is an equivalent replacement for the likelihood. Given each forecast category, the number of prospects when the result is dry, is plotted against the number of prospects when the result is a discovery. The discrimination diagram is constructed by plotting the number of prospects for the dichotomous results for all the forecasting categories.

The discrimination diagram for 3 sets are in fig. 5.8. There are little overlaps for forecasts less than 20%, for all three sets, so good discrimination for forecasts less than 20%.Much overlap for forecast probabilities larger than 30%, so poor discrimination for forecasts larger than 30% that the forecasts does not discriminate discovery well from dry.

5.4 ROC analysis

introduction

Receiver operating characteristic (ROC) analysis was originally developed in signal detection theory as a model for how to separate the signal from the Gaussian noise. Now ROC analysis has been widely used in medical diagnosis tests and has been increasingly used in other fields such as data mining and atmospheric science. ROC framework is a bit different from the Murphy & Winkler framework which discourages the reduction of the forecasts into categories (Marzban, 2003). The ROC analysis is based on contingency table, which categorize the probability forecasts into bins. ROC analysis is still within the Murphy & Winkler framework.

ROC curve is a two dimensional measure of classification performance, i.e., measuring the decision threshold introduced to produce binary classifications. It has potential value as it can indicate weather to drill based on the chance of geologic success. The area under the ROC curve (AUC) is a scalar measure of one facet of the forecast performance. ROC graphs are a very useful tool for visualizing and evaluating classifiers- the threshold to make binary descion. They are able to provide a richer measure of classification performance (Fawcett, 2006).

Contingency Table

		Observed: yes	Observed: no	Total
Forecast	yes	*hits*	*false alarms*	*forecast yes*
	no	*misses*	*correct negatives*	*forecast no*
Total		*observed yes*	*observed no*	*total*

Figure 5.9: contigency table for binary forecasts (CAWCR, 2014)

Binary Classification

ROC Curve is based on *hitrates* and *false alarm ratio* calculated from contingency table for binary forecasts as fig. 5.9. By introducing decision threshold, probability forecasts can be transformed to binary classification, "A classifier need not produce accurate, calibrated probability estimates; it need only produce relative accurate scores that serve to discriminate positive and negative instances" (Fawcett, 2006).

- **Hit rate (H)**: $H = \frac{hits}{hits+misses}$

  Range: 0 to1 , Perfect score: 1.

  H is sensitive to hits, but ignores false alarms. Estimate of $p(f = 1|x = 1)$

- **False alarm ratio (F)**: $F = \frac{false_alarms}{hits+false_alarms}$

  Range: 0 to1, Perfect score: 0

  Sensitive to false alarms, but ignores misses. Estimate of $p(f = 1|x = 0)$

- **Peirce skill score (PSS)**: $PSS = H - F$

  Range -1 to1, Perfect score: 1

  PSS is a measure of skill obtained by the difference between the hit rate and the false alarm rate. If the PSS is greater than zero, then the number of hits exceeds the false alarms and the forecast has some skill.

Decision threshold

For probability predictions, suppose there is a decision variable W, W exceeding a probability forecast threshold w can be interpreted as signal for occurrence; the non-occurrence is otherwise. The decision threshold variable can function as a binary classifier. In our case, the threshold(w) is the chance of geologic success that decision would be made to whether to predict the prospect at eitehr success or fail. So the probability are transformed into binary forecasts. Then the hit rate and false alarm ratio are defined:

$$H(w) = Pr(W \geq w)|X = 1) \tag{5.1}$$

$$F(w) = Pr(W \leq w)|X = o) \tag{5.2}$$

ROC curve

The ROC curve is then obtained by plotting the calculated hit rate and false alarm rate for all probability categories.

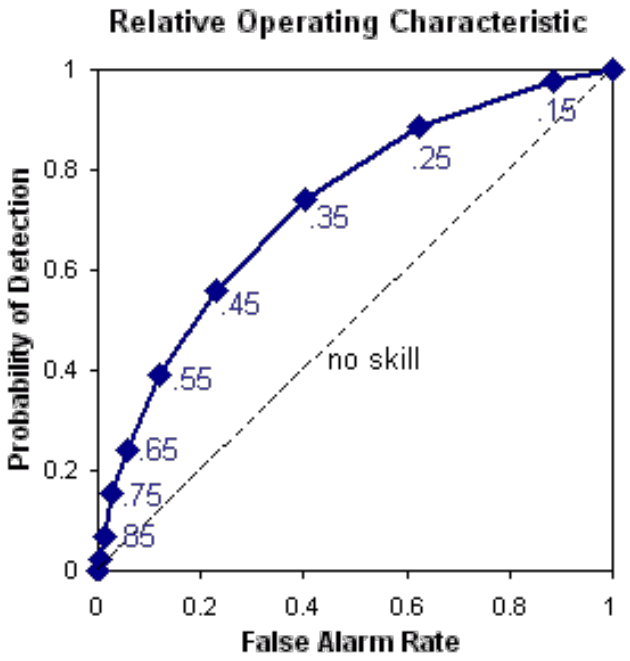


Figure 5.10: ROC Curve (CAWCR, 2014)

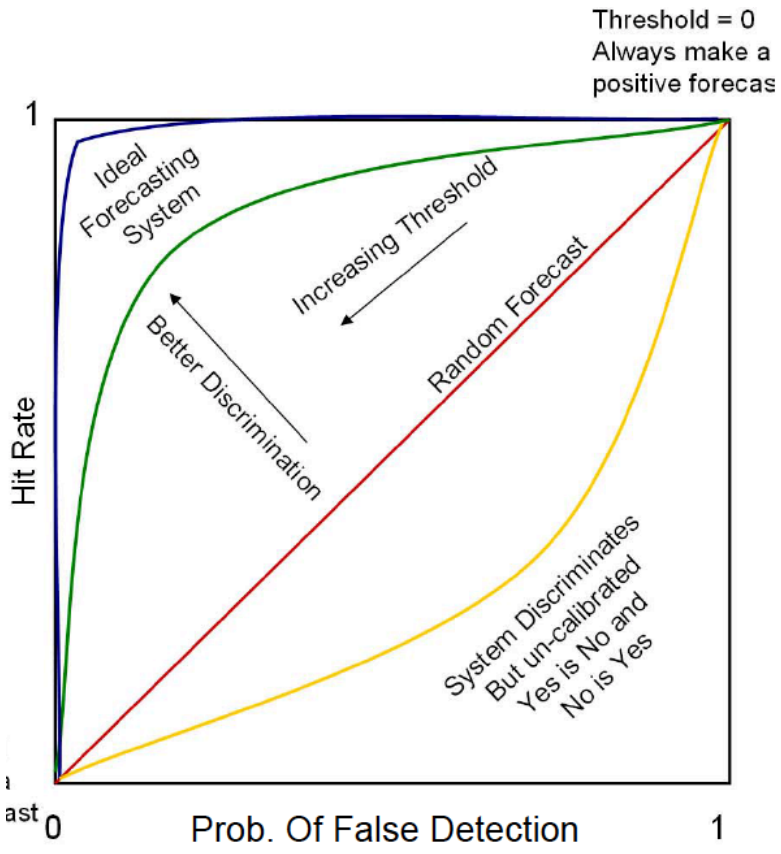


Figure 5.11: Interpreation of ROC curve (CAWCR, 2014)

An ROC curve is a two-dimensional depiction of classifier performance including H and F. The ROC is conditioned on the observations (i.e., given the outcome, what was the corresponding forecast. ROC would be good company to the reliability diagram, which is conditioned on the forecasts. The diagonal line is *no skill line* on ROC curve, show in fig. 5.10. Perfect ROC curve travel from top right of diagram, to top left, then across to bottom left of diagram, as probability thresholds increases. Low thresholds lead to both high H and F towards the upper right hand corner; High thresholds make the ROC points move towards the lower left corner. An interpretation of the ROC curve is in fig. 5.11.

Area under ROC (AUC) is often used as a score measuring skill of forecast. AUC gives one single-number measurement of performance. AUC aggregates performance across the entire range of probabilities.

Range: 0 to 1; 0.5 indicates no skill.

Perfect score: 1

Applied to 3 sets of data

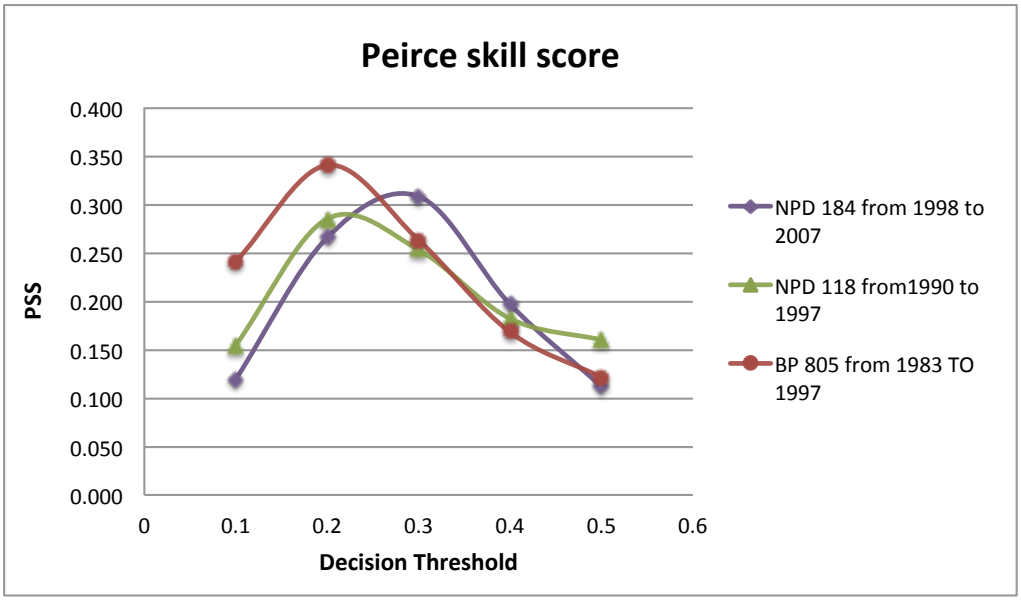


Figure 5.12: PSS for 3 sets of data

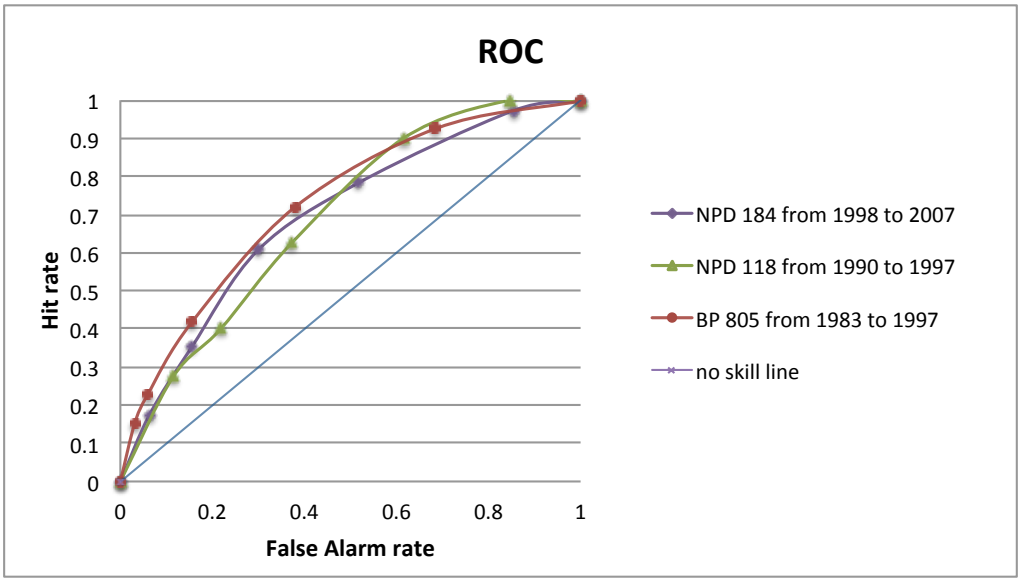


Figure 5.13: ROC curve for 3 sets of data

PSS measures the difference between hit rate and false alarm rate. PSS thus indicate how well a specific chance of geologic success, as a threshold for binary prediction, separates Hits/success from false alarm ratio/failure. In fig. 5.12, for example, the PSS value is highest at probability around 30% for NPD 184 dataset. When using 30% chance of geologic success as the thresh-

Decision Threshold w	F(f>=w|x=0)	H (f>=w|x=1)	PSS
NPD 184 from 1998 to 2007			
0	1	1	0
0.1	0.855	0.973	0.118
0.2	0.518	0.784	0.266
0.3	0.300	0.608	0.308
0.4	0.155	0.351	0.197
0.5	0.064	0.176	0.112
NPD 118 from1990 to 1997			
0	1	1	0
0.1	0.846	1.000	0.154
0.2	0.615	0.900	0.285
0.3	0.372	0.625	0.253
0.4	0.218	0.400	0.182
0.5	0.115	0.275	0.160
BP 805 from 1983 to 1997			
0	1	1	0.000
0.1	0.686	0.926	0.240
0.2	0.380	0.721	0.340
0.3	0.156	0.419	0.263
0.4	0.059	0.227	0.168
0.5	0.031	0.153	0.122

Table 5.1: PSS, Hit rate and False alarm rate results for 3 datasets

old to decide whether the prospect is dry or success, most skill are reached since this threshold mostly separates a success from dry.

In ROC fig. 5.13, BPs 805 has largest area under the ROC curve and thus best classification performance. In other words, the forecast of BP are best in discriminating discovery from dry among the three sets of data.

I

Chapter 6

Handling Small Size Data

One limiting factor for chance of geologic success verification is that the number of pairs of prognosis and observation data available is ususally small compared to data size in climatology. For instance, it would be very likely that less than 100 pairs of data is available, when an oil company want to evaluate its geologic success prognosis performance in a region for a 5 year period. The DO approach outlined by Murphy and Winkler (1987) is based on joint, marginal and conditional distribution of forecasts and observations, which requires binning the discrete probabilistic forecasts into several categories. When the data is limited, the data in some bin would be very few ,leading to distorted estimation.

The difference betweenn meassure acquired by continuous

Dimensionality(D) is the number of degrees of freedom in order to estimate the joint distribution of forecasts and observations. For instance, there are n_f categories of prognosis, and n_x cases of observations, the dimensionality is defined as (Hashino et al., 2002) $D = n_f * n_x - 1$. For example, when geologic success probability are categorized into 6 bins and observations are binary: discovery and dry/small. $D = 6*2-1 = 11$, which means that at least 11 pairs of data are necessary to estimate the relevant measures for *contigeny talbe method*. One would expect that very few data less than 11 in some forecast categories if the total datasets are around 50 or less than 100.

There are differecnces between verification measures aquired from continuous approach described in chapter 4 section 4.3 and contigency table method. The smaller size of the data, the more diffrence would be observed. The contigency table method woould distort much for small

size data, as Dimensinoalty is too high. In continuous approach, though Most scalar measrues for verifcation can be direcetly estimated from original sample data by analytical expressions, $\mu_{x|f}$ in calibration refinement factorization measures needs to be estimated by alternative approaches. Statistical modeling techniques suggested by Bradley et al. (2003b) can be useful to estimate CR measures when data size is small. Bradley et al. (2003b) did Monte Carlo experiments for forecasting examples, and the results show that use of statisctical modeling approach significantly improve the estimated of the CR measures for pairs of data of 500 or less. Recommended statistical methods are applied to the data of 184 prospects on NCS. Logistic regression method is used to estimate the mean conditional distribution. And Kernel density estimation is used to estimate the distribution of the probability of discovery. The two methods are presented below.

6.1 Logistic Regression (LR)

Logistic regression is a regression model where the dependant variable is categorical. The binary logistic model is used to estimate the probability of a binary response. The logistic regression model could estimate the mean $\mu_{x|f}$.

$$Logit(P) = \log \frac{P}{1-P} = \beta_0 + \beta_1 * f \tag{6.1}$$

The logistic regression of mean P for probability forecast f, is expressed with two parameters β_0 and β_1 as in equatoin 6.1. After we get the two paramters. The estimator of the conditional mean could be calculated from equation 6.2:

$$\mu_{x|f_i} = Pr(X_i = 1) = \frac{\exp(\beta_0 + \beta_1 * f_i)}{1 + \exp(\beta_0 + \beta_1 * f_i)} \tag{6.2}$$

The two paremeters β_0 and β_1 can be estimated by method of maximum likelihood or by generalized linear model (GLM).

The LR is realized by **Generalized linear model** here, since GLM is simpler to implement. Logistic regression can be seen as a special case of Generalized linear model and thus are analgous to a linear model. A Link function allow the GLM to fit the logistic regression model. For

binomial data, a logit link, as equation 6.3, is mostly used.

$$Logit(P) = \ln \frac{P}{1-P} \tag{6.3}$$

The linear function of the predictor variables is calculated and the result of this calculation is run through the link function 6.3. The *glmfit* function in Matlab is applied to model the regression. it returns -1.69 and 4.26 for the value of β_0 and β_1 in matlab. Continuous $\mu_{x|f_i}$ can be obtained now and is plotted in fig 6.1.

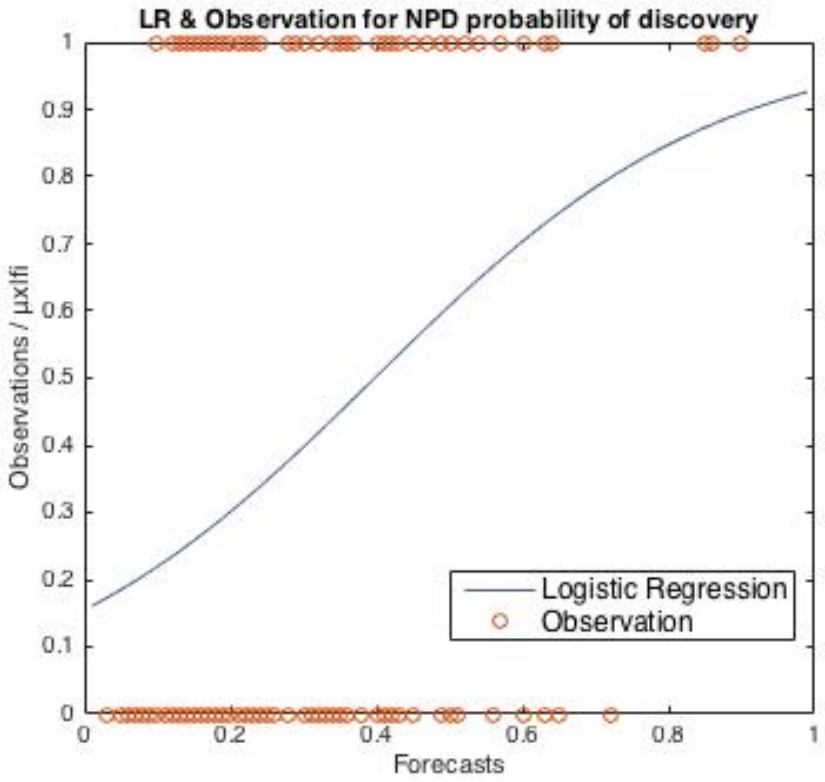


Figure 6.1: $\mu_{x|f_i}$ obtained by logsitic regression method

6.2 Kernel Density Estimation (KDE)

To visualize this one dimensional distribution of prediction, histogram in the sharpness diagram is one way, and KDE is another way that can represent the continuous distribution of the forecasts. KDE is a non-parametric method to estimate the probability density function. a *ker-*

nel function is a non-negative function that integrates to one and has mean zero. Basically, KDE is usieng a kernel fucntion to smooth (making inferences) aournd given observed data points. The baisc kerndel denstiy estimator is 6.4:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - x_i}{h}) \tag{6.4}$$

h denotes the bandwidth; n is the number of samples, and $k()$ is the kernel funciton. The big-weight kernel 6.5 is utilized:

$$k(t) = \frac{15}{16}(1 - t^2)^2 \tag{6.5}$$

The larger the bandwidth, the more influences the observed datappoints have on the KDE curve. "The bandwitdth is often estimated from sample data by cross validation or other approaches that attempt to find the best fit to the data" (Bradley et al., 2003b). An optimum bandwidth deterimed through the normal referecne rule 6.6 described by Bradley et al. (2003b). The default bandwith in *Matlab* for optimal normal density returns 0.0609, higher than the bandwdith,0.0307 calcluated by the noraml referecne rule 6.6. Distributions for both bandwith are plotted in fig. 6.2.

$$h = 2.623(4/3)^{1/5}\sigma N^{-1/5} \tag{6.6}$$

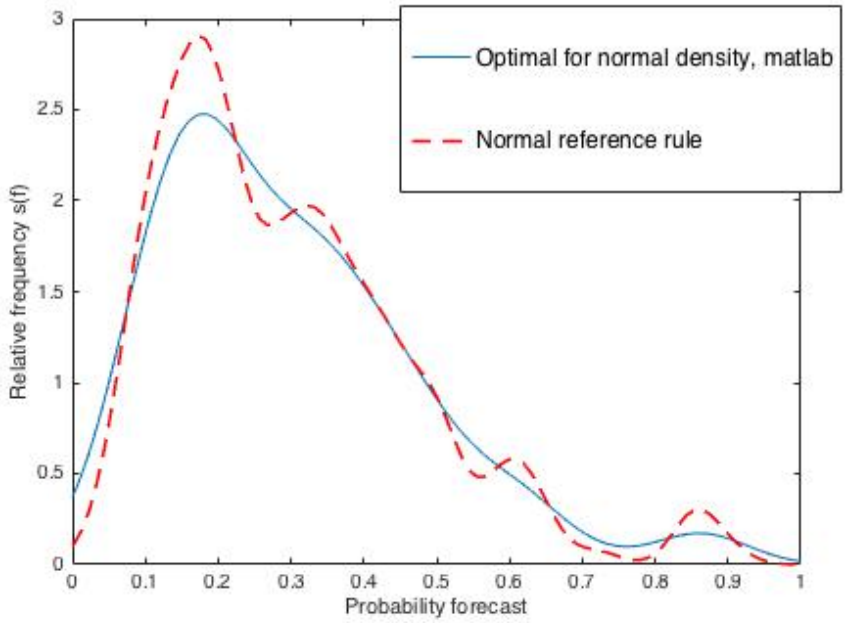


Figure 6.2: Distribution of the probability of success for 184 prospects on NCS by KDE

Chapter 7

Summary and Discussion

7.1 Chance of Geologic Success and Inconsistency

Literatures show that there is inconsistency in comparing post-drill result with pre-drill forecast; for example, geologic success was compared with discovery, where discrepancy can exist between the two concepts. The definition of geologic success has been explicitly given by Rose (2001), but still needs adaptations in different cases, as there may not be corresponding post-drill record that completely matches the exact geologic success. Usually, the definition of geologic success is adapted to involve some minimum volumetric of hydrocarbons, which would then correspond to documented records. The probability of discovery is commonly obtained by multiplying probability of geologic chance factors. There is also alternative way where historical success rate is taken into account. Either way, the evaluation comes down to evaluating geologic chance factors, or geologic sub-factors where explorers place their judgement in probabilistic form utilizing their experience, knowledge, information acquired from geotechnical and geologic measures. To ensure consistency and accuracy, tools as *chance adequacy matrix* and *risk table* for specific geologic factor by which probability values or intervals are corresponded to detailed qualitative description of different dimensions including geologic knowledge, confidence or certainty level, and data quality.

It is still hard to overcome bias and heuristics when making subjective judgment. Subjective probability are being criticized for its inconsistency. Geoscientists are also susceptible to common individual bias. Chance of geologic success is usually assigned by a group of experts.

Thus, more complicated group bias caused by group interactions further influences the probability assessment. Elicitation methods are available to handle both individual and group bias, but elicitation usually involves several steps and takes a long time span, and verification can be part of the elicitation process.

7.2 Verification

By assigning subjective probabilities, however, experts' knowledge and experiences are being utilized, and uncertainty or lack of information are properly conveyed to others. Extensive statistical verification measures have been extensively developed in atmospheric science for several purposes including improving the forecast performance. To improve probability assigning performance, verification measures to study post-drill result and pre-drill predictions should be explored.

The verification measures employed in exploration ventures found on literatures are mostly rudimentary and even erroneous. Extensive verification measure are available and thus pertinent measures for probability predictions are presented and applied to data. Summary measures provides a single valur or score for evaluating the overall prognosis quality or performance. Distribution-oriented approaches provides more detailed information in terms of measures as reliability, discrimination, and resolution on each prediction interval. Thus performance can be compared between different probability interval for the same dataset. Different characteristics in different interval of the dataset applied are clearly demonstrated through these DO verification measures Measures calculated by continuous approaches also provides exact numerical values.

It is not easy to grasp useful information from these scalar measure in the beginning. These graphical measures provides more direct information. prediction qualities as reliability, sharpness, discrimination and resolution are demonstrated in graphs. One can read these qualities in specific probability interval directly from the graph and compare them with other datasets graphically as measures are applied to 3 datasets.

ROC is an excellent tool for assessing the binary classification. the Area under the ROC curve is used as a performance score, considering forecasts' ability to discriminate between Yes and

No events while calibration is not considered by. ROC analysis including Hit rate, false alarm rate and PSS give performance measurement to individual decision thresholds, which may be helpful for decision-makings based on chance of geological success. The presentation of PSS is more exploratory. The realistic value of PSS to chance of geologic success remains to be further studied. Though ROC is commonly used, it is more suitable for rare events and the usefulness in helping drilling decision needs to be further studied.

When there are enough data, graphical measures are good as the curves are more smooth. For smaller size data, the curve looks zigzag and measures calculated from contingency table methods could provide distorted information about the forecast quality. statistical methods as logistic regression and kernel density method can significantly improve the estimates of calibration-refinement measures which can not be directly obtained from oringinal datasets.

Not all measures may be of interests to the assessors. On the other hand, not a single attribute or measure would provide a whole picture of the forecast situations. It remains to be further studied to whether some measures are pertinent to chance of geologic success. Many measures may seem to be comprehensive at first hand. As long as you have a understanding of verification measures, they provide intuitive information about strength and weakness of the predictions in specific situations. The verification measures provide more formalized and useful feedback, so that explorers are more encouraged to give more ”true” probability corresponding to their judgement. Their performance has a more ”tangible” record. And ultimately, their probability calibration performance can be improved.

Appendix A

Appendix

Risk factors considered in different methodologies of estimating the total geological probability of success (PoS).

Authors	Company	Number of risk factors	List of geological risk factors	How total geological PoS was calculated
White (1993)		12 (4 groups)	Closure volume, seal, timing, reservoir facies thickness, porosity, permeability and continuity, organic quantity and quality, maturation, migration, preservation, hydrocarbon quality and concentration, recovery.	Serial multiplication of 12 factors.
Goldstein (1994)	Bridge Oil	33 (5 groups)	Trap, seal, source, reservoir, likelihood of oil.	Least favorable of all factors in each group was carried into final calculations ("weakest-link" logic). Serial multiplication of 5 factors.
Duff and Hall (1996)	PetroFina	4	Reservoir, seal, hydrocarbon charge, closure.	Serial multiplication of 4 factors.
Hermanrud et al. (1996)	Statoil	7	Closure, reservoir, porosity, source/migration, timing, trap, recovery.	Serial multiplication of 7 factors.
Snow et al. (1996)	Conoco	4	Reservoir, trap, seal, source.	Used historical success rate for the play and then multiplied that success rate by four comparison coefficients (0.5 to 1.5). Then, serial multiplication of 4 risk factors.
Otis and Schneidermann (1997)	Chevron	14 (4 groups)	Presence of mature source rock, presence of reservoir rock, presence of a trap, and play dynamics (timing).	Least favorable of all factors in each group was carried into final calculations ("weakest-link" logic). Serial multiplication of 4 factors.
Alexander and Lohr (1998)	Unocal	5	Trap, top seal, reservoir rock, source rock, timing/migration.	Serial multiplication of 5 factors.
McMaster (1997)	Amoco	6	Trap (structure), seal, reservoir, porosity, source, migration.	Serial multiplication of 6 factors.
Johns et al. (1998)	Santos	4	Closure, reservoir, seal, charge.	Serial multiplication of 4 factors.
Watson (1998)	BHP Petroleum	50 (7 groups)	Trap geometry, source, migration and timing, seal, reservoir, preservation, gas risk, DHIs.	Combined according to specified rules.
CCOP (2000)		7	Reservoir facies presence, reservoir effectiveness, trap presence/definition, seal effectiveness, presence of mature source rock, migration effectiveness, retention.	Serial multiplication of 7 factors.
Wang et al. (2000)	Texaco	4	Structure, seal, reservoir and charge.	Serial multiplication of 4 factors.
Rose (2001a)		15 (5 groups)	Source rocks, timing/migration, reservoir rock, closure, containment.	Least favorable of all factors in each group was carried into final calculations ("weakest-link" logic). Serial multiplication of 5 factors.
Lowry et al. (2005)	Origin	19 (5 groups)	Closure, reservoir, source, charge and seal	Serial multiplication of 19 factors.
Sykes et al. (2011)	ExxonMobil	9	Trap closure, trap seal, reservoir facies presence, reservoir quality, source richness, source maturation, migration pathways, trap-migration timing, hydrocarbon recovery.	Serial multiplication of 9 factors.

Figure A.1: Risk factors considered in different methodologies of estimating the total geological probability of success (Milkov, 2015)

Table for sets of general qualitative descriptions for the relative probability scale from *The CCOP Guidelines for Risk Assessment of Petroleum Prospects* (CCOP, 2000).

Risk table for estimating probability of structure (closure, geometry, container) being present

(Milkov, 2015)

P	General scale	Analogue or theoretical models	Proven geological models	P
1.0	Condition is **virtually to absolutely certain**. Data quality and control is excellent.	Only possible model applicable for the concerned area. Unfavourable models are impossible.	Identical geological factor to those found in fields and discoveries in immediate vicinity. Conditions are verified by unambiguous well and seismic control.	1.0
0.9		The model is very likely to absolutely certain. Unfavourable models are not impossible.		0.9
0.8	Condition is **most probable**. Data control and quality is **good**. Most likely interpretation.	The model is very likely. Only minor chance that unfavourable models can be applied.	Similar geological factor successfully tested by wells in the trend. Lateral continuity is probable as indicated by convincing well and seismic control.	0.8
0.7		The model is likely to very likely. Unfavourable models can be applied.		0.7
0.6	Condition is **probable** or data control and quality is **fair**. Favourable interpretation.	The model is mor likely than all otherunfavourable models.	Similar geological factor is known to exist within the trend. Lateral continuity is probable as indicated by limited well and seismic data.	0.6
0.5		Likely model, however, unfavourable are also likely.		0.5
0.4	Condition is **possible** or data control and quality is **poor to fair**. Less favourable interpretation possible.	Unfavourable models are more likely than applied model.	Similar geological factor may exist within the trend. Valid concepts, but unconvincing data only hints at possible presence of the feature.	0.4
0.3		The model is questionable. and unfavourable models are likely to very likely.		0.3
0.2	Condition is **virtually to absolutely impossible**. Data control and quality is **excellent**.	The model is unlikely and very questionable. Unfavourable models are very likely.	The geological factor is not known to exist within the trend. Conditions are verified by unambiguous well and seismic control.	0.2
0.0		The model is unlikely and highly questionable. Unfavourable models are very likely to certain.		0.0

Figure A.2: Qualitative descriptions for the relative probability scale (CCOP, 2000)

Structure (closure, geometry, container)				Data (existence and reliability)			
				3D seismic	2D seismic Number of lines per structure (with obligatory availability of in-line and cross lines)		
					Dense (7 lines and more)	Sparse (3-6 lines)	Very sparse (2 lines) (Lead)
Models (existence and reliability)	Seismic mapping and correlation	High-relief structure (≥3 times higher than seismic accuracy) AND low structural complexity (4-way)	Easy to interpret, reliable correlation based on nearby (<50 km) wells	1.00	0.90	0.80	0.60
			Uncertain correlation (horizons are interrupted laterally) or based on remote (> 50 km) wells	0.95	0.85	0.75	0.55
			Difficult to interpret, unreliable correlation (horizons are interrupted by thrust faults, diapirs, etc.) or model developed using analogues without wells in the basin	0.85	0.75	0.70	0.45
		Medium-relief structure (1-3 times higher than seismic accuracy) OR high-relief structure with high structural complexity (3-way, stratigraphic)	Easy to interpret, reliable correlation based on nearby (<50 km) wells	0.80	0.70	0.60	0.35
			Uncertain correlation (horizons are interrupted laterally) or based on remote (> 50 km) wells	0.75	0.65	0.50	0.25
			Difficult to interpret, unreliable correlation (horizons are interrupted by thrust faults, diapirs, etc.) or model developed using analogues without wells in the basin	0.70	0.55	0.45	0.20
		Low-relief structure (lower than seismic accuracy) OR high uncertainty of depth conversion (subsalt, below lava flows) OR areas with rapidly changing lateral velocities in the overburden	Easy to interpret, reliable correlation based on nearby (<50 km) wells	0.55	0.45	0.35	0.15
			Uncertain correlation (horizons are interrupted laterally) or based on remote (> 50 km) wells	0.50	0.40	0.25	0.10
			Difficult to interpret, unreliable correlation (horizons are interrupted by thrust faults, diapirs, etc.) or model developed using analogues without wells in the basin	0.40	0.30	0.20	0.05
		Low-relief structure (lower than seismic accuracy) AND EITHER high uncertainty of depth conversion (subsalt, below lava flows) OR areas with rapidly changing lateral velocities in the overburden		0.35	0.25	0.15	0.05

Figure A.3: Qualitative descriptions for the relative probability scale (Milkov, 2015)

Bibliography

Baddeley, M. C., Curtis, A., and Wood, R. (2004). An introduction to prior information derived from probabilistic judgements: elicitation of knowledge, cognitive bias and herding. *Geological Society, London, Special Publications*, 239(1):15–27.

Bradley, A. A., Hashino, T., and Schwartz, S. S. (2003a). Distributions-oriented verification of probability forecasts for small data samples. *Weather and forecasting*, 18(5):903–917.

Bradley, A. A., Hashino, T., and Schwartz, S. S. (2003b). Distributions-oriented verification of probability forecasts for small data samples. *Weather and forecasting*, 18(5):903–917.

Bratvold, R. and Begg, S. (2010). *Making good decisions*. Society of Petroleum Engineers.

CAWCR (2014). Forecast verification: issues, methods and faq.

CCOP (2000). The ccop guidelines for risk assessment of petroleum prospects.

Curtis, A. and Wood, R. (2004). Optimal elicitation of probabilistic information from experts. *Geological Society, London, Special Publications*, 239(1):127–145.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701.

Gofa(HNMS) (2010). Guidelines for verification of ensemble forecasts.

Harbaugh, J. W., Davis, J. C., and Wendebourg, J. (1995). *Computing Risk for Oil Prospects: Principles and Programs: Principles and Programs*. Elsevier.

Harper, F. G. (2000). Prediction accuracy in petroleum prospect assessment: A 15 year retrospective in bp. *Norwegian Petroleum Society Special Publications*, 9:15–21.

Hashino, T., Bradley, A. A., and Schwartz, S. S. (2002). *Verification of probabilistic streamflow forecasts*. PhD thesis, University of Iowa.

Hora, S. (2007). Eliciting probabilities from experts. *Advances in Decision Analysis: From Foundations to Applications*, page 129.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Kahneman, D. and Tversky, A. (1982). Variants of uncertainty. *Cognition*, 11(2):143–157.

Kari Ofstad, P. B. and Helliksen, D. (2015). Prognoses and results of wildcat wells drilled between 1998 and 2007 on the norwegian continental shelf.

Magoon, L. B. and Dow, W. G. (1994). The petroleum system. *The petroleum system—From source to trap: AAPG Memoir*, 60:3–24.

Marzban, C. (2003). A comment on the roc curve and the area under it as performance measures. *Wea. Forecasting*.

Milkov, A. V. (2015). Risk tables for less biased and more consistent estimation of probability of geological success (pos) for segments with conventional oil and gas prospective resources. *Earth-Science Reviews*, 150:453–476.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338.

NPD (2010). Prospects and plays.

Ofstad, K., Kullerud, L., and Helliksen, D. (2000). Evaluation of norwegian wildcat wells (article 1). *Norwegian Petroleum Society Special Publications*, 9:23–31.

Rose, P. R. (1992). Chance of success and its use in petroleum exploration: Chapter 7: Part ii. nature of the business.

Rose, P. R. (2001). *Risk analysis and management of petroleum exploration ventures*, volume 12. American Association of Petroleum Geologists Tulsa, OK.

Snow, J., Dore, A., and Dorn-Lopez, D. (1996). Risk analysis and full-cycle probabilistic modelling of prospects: a prototype system developed for the norwegian shelf. *Norwegian Petroleum Society Special Publications*, 6:153–165.

Welsh, M., Lee, M., and Begg, S. (2008). More-or-less elicitation (mole): Testing a heuristic elicitation method. In *Annual Meeting of the Cognitive Science Society (30th: 2008: Washington DC)*.

Winkler, R. L. and Murphy, A. H. (1968). "good" probability assessors. *Journal of applied Meteorology*, 7(5):751–758.