Spatial Dependency, Aggregation and Data

by
Jack Schuenemeyer
Southwest Statistical Consulting, LLC
Cortez, Colorado USA

2012 International Association for Mathematical Geosciences Distinguished Lecturer

Indian Statistical Institute

Acknowledgements

- International Association for Mathematical Geosciences for financial support
- Professor Sagar, ISI for very kind hospitality here in Bangalore
- Some work is joint with
  - Dr. Donald Gautier, US Geological Survey
  - Dr. Gordon Kaufman, Professor Emeritus, MIT

International Association for Mathematical Geosciences Student Chapters

- Student Chapters
  - You can plan events
  - Interact with other students
  - Receive financial help attending conferences
  - Meet practicing scientists
- Earth, climate, and environmental sciences are exciting disciplines, which combined with mathematics and statistics, can help solve important problems that will benefit us and future generations

Outline

I. Dependency & Aggregation

II. Data

   1. Hard
   2. Analogs
   3. Expert Judgment

Selected Spatial Earth Science Applications

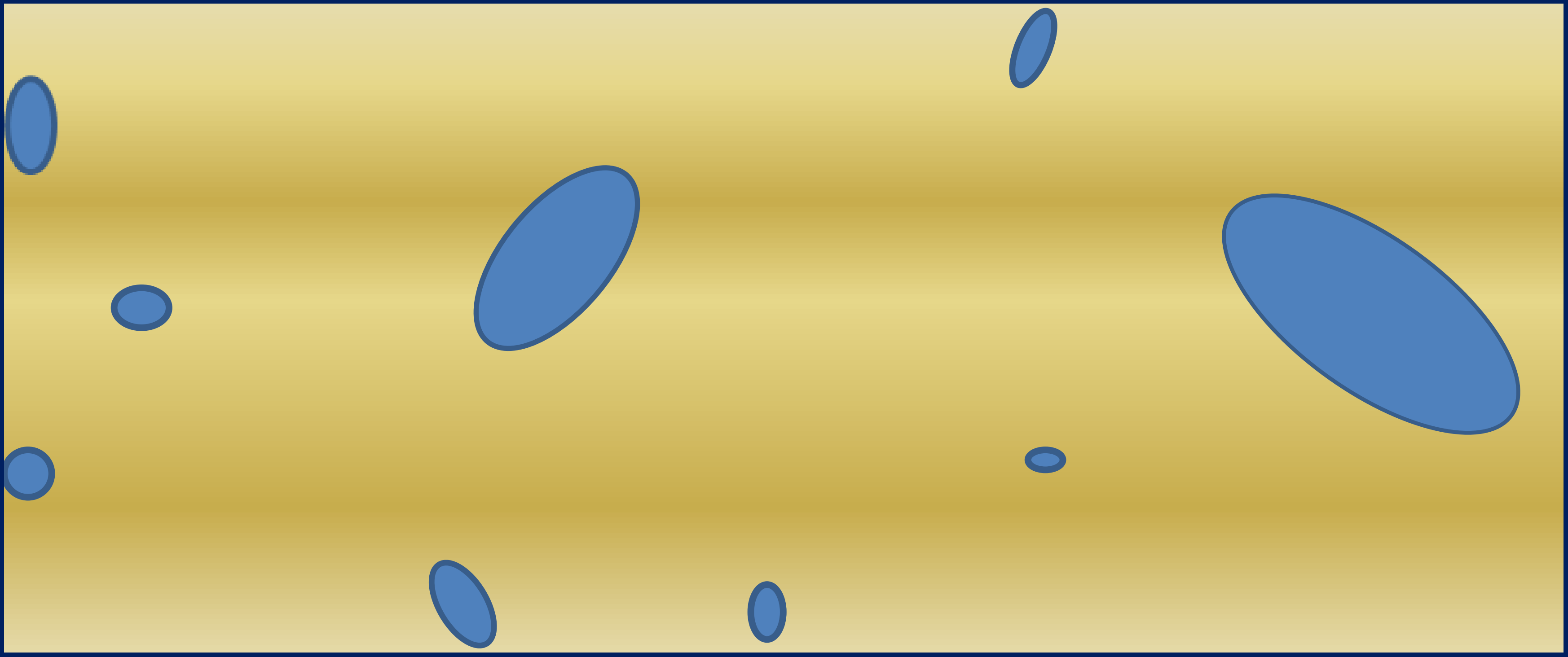
- Estimating remaining mineral or energy resources
- Identifying characteristics of a specific resource (energy companies)
- Modeling geologic hazards
- Transport systems – water, hydrology
- Snow melt

I. Dependency & Aggregation

Estimating Remaining Exhaustible Energy Resources

- Important for governments, energy companies, research institutes
- Oil and gas occur in relatively well defined regions called basins or plays
- Two types of oil and gas resource
  - Conventional ⟷ Discrete
  - Continuous
- Most regions at least partially explored

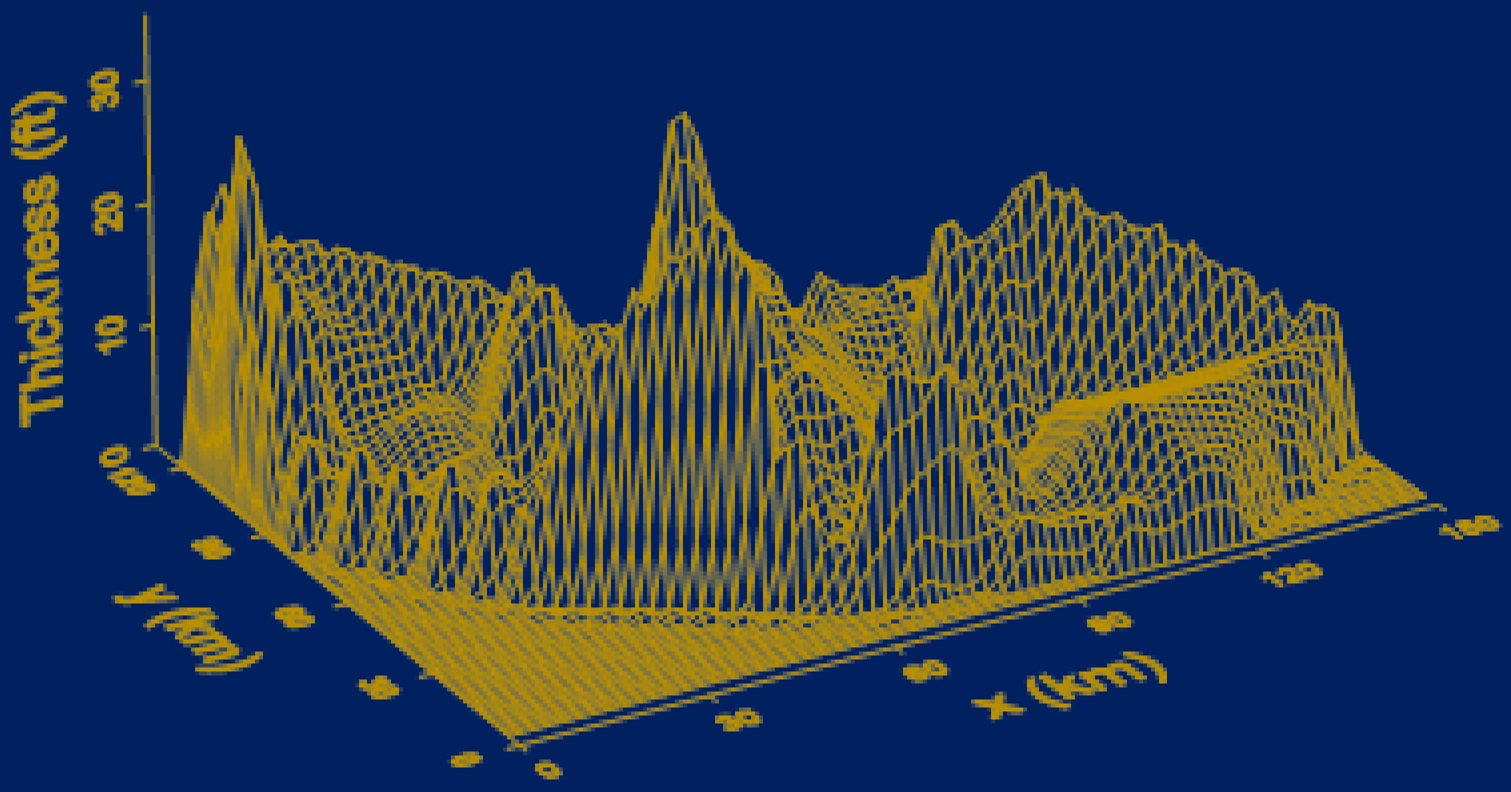
Conventional (discrete)



Tasks

- Understanding and modeling the discovery process
  - Point process. Remaining resource function of discovered resource and efficiency of sampling
- Temporal component is modeling the learning curve in the discovery process

Continuous



Issues

- Concentration of resources varies continuously – estimate regions of higher concentration

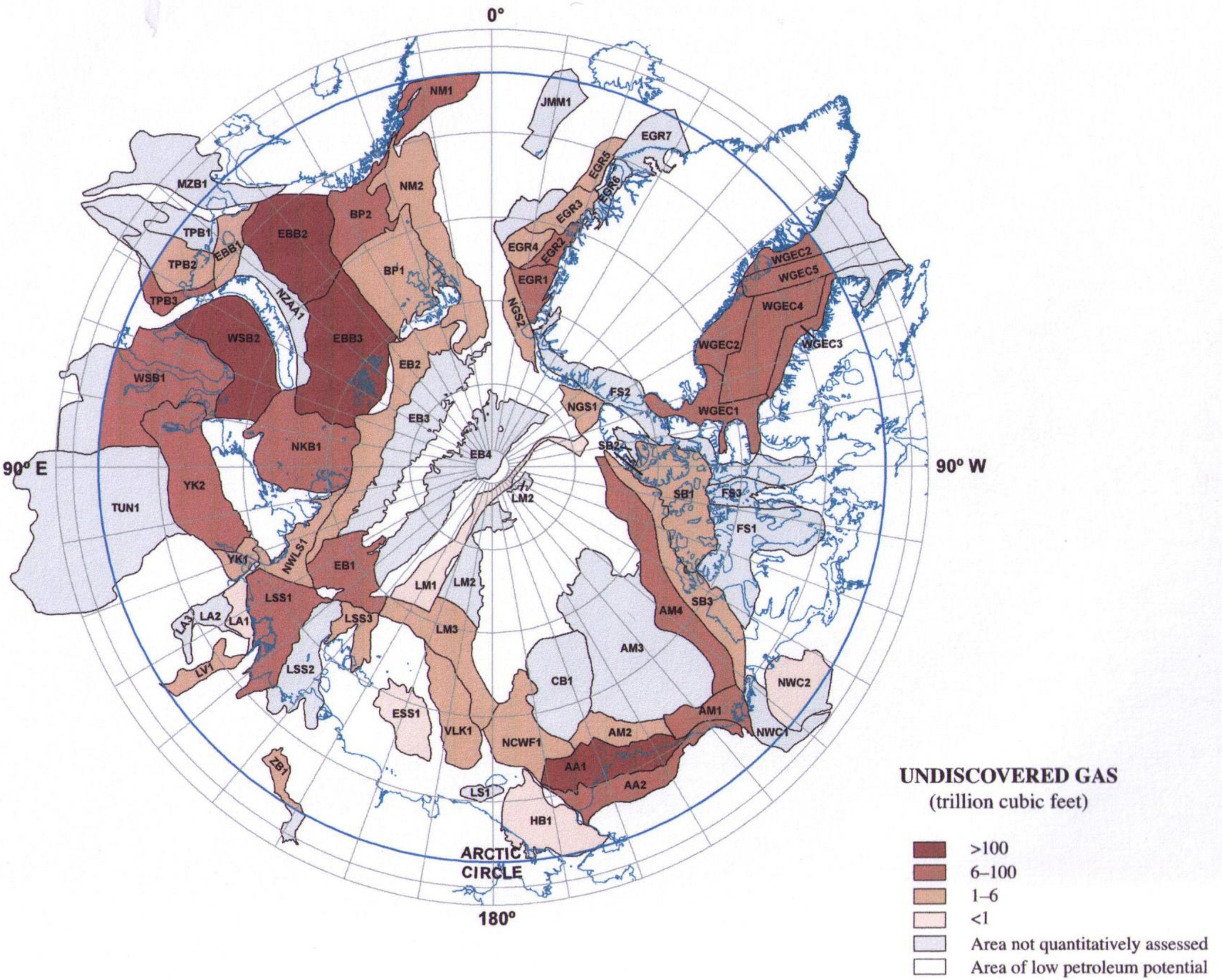
- Spatial model – fit to partial data
- Cell based
- Nearest neighbor
- Clearly there are spatial dependencies!

Aggregation of Results

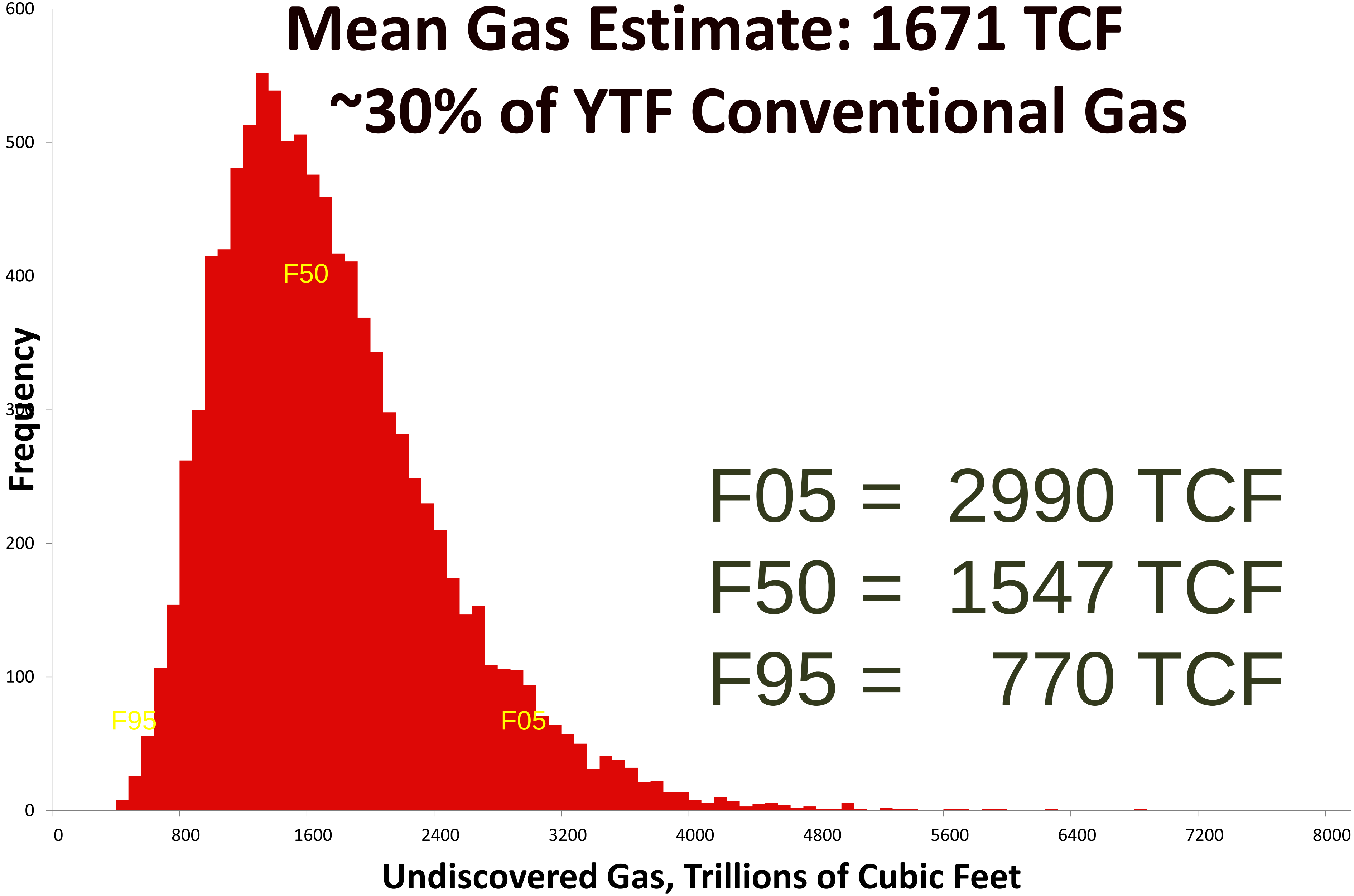
- Interest to governments
- Large energy companies
- Tax policy
- Research institutes

Assessing Dependency

- Consider 2 adjoining regions A and B.
- If the assessor overestimate resources in A does that mean that imply that resources in region B were over estimated?
- If answer is YES regions A and B are dependent
- If answer is NO regions A and B are independent


0°
90° E
90° W
180°
ARCTIC CIRCLE
NM1
JMM1
EGR7
EGR5
EGR6
NM2
MZB1
BP2
EGR3
EBB2
TPB1
EBB1
EGR4
EGR2
BP1
WGEC2
WGEC5
TPB2
EGR1
NZAA1
TPB3
NGS2
WGEC4
WSB2
EBB3
WGEC2
WGEC3
EB2
WSB1
FS2
NGS1
WGEC1
EB3
NKB1
SB2
EB4
90° E
YK2
SB1
FS3
TUN1
LM2
FS1
NWLS1
YK1
EB1
LM1
LM2
LSS1
SB3
AM4
LA3
LA2
LA1
LSS3
LM3
AM3
LV1
LSS2
CB1
NWC2
ESS1
AM1
VLK1
AM2
NWC1
NCWF1
ZB1
AA1
AA2
LS1
HB1
UNDISCOVERED GAS
(trillion cubic feet)
>100
6–100
1–6
<1
Area not quantitatively assessed
Area of low petroleum potential

Mean Gas Estimate: 1671 TCF
~30% of YTF Conventional Gas



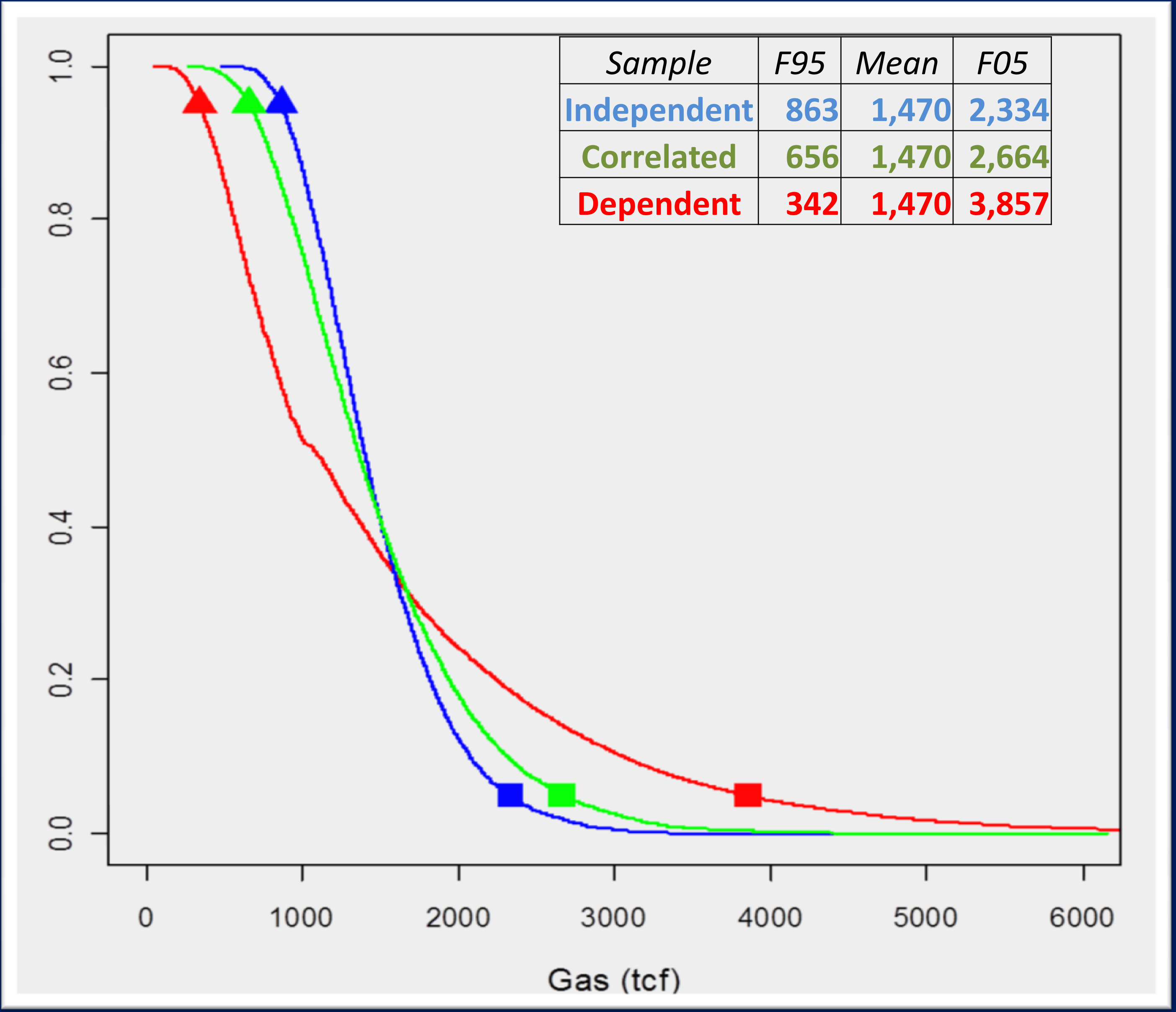
F05 = 2990 TCF

F50 = 1547 TCF

F95 = 770 TCF

Aggregated Gas – Circum Arctic

Inverse
cdf



Sample	*F95*	*Mean*	*F05*
Independent	**863**	**1,470**	**2,334**
Correlated	**656**	**1,470**	**2,664**
Dependent	**342**	**1,470**	**3,857**

Types of Dependencies

- Physical
  - Attributes correlated
- Human
  - Same assessment team
  - Same organization

Implications of Dependency Assumptions

- Effect on aggregated results
  - Pairwise independent – uncertainty *too small*
  - Totally dependent – uncertainty *too large*
- NEITHER ASSUMPTION USUALLY VALID

Correlation vs. Dependency

- Correlation is one measure of dependency
  - Many measures of correlation
- Correlation is not affected by parameter changes (size and/or shape of oil or gas distributions)
- Dependency can be modeled via regression

Independent

AU 1 Gas (tcf)	AU 2 Gas (tcf)
11.5	9.1
6.7	8.4
5.5	2.1
0.1	3.8
13.7	0.2
11.7	1.4
10.9	7.6
0.3	3.7
7.6	9.9
26.2	1.9

Fractile Dependent

AU 1 Gas (tcf)	AU 2 Gas (tcf)
0.1	0.2
0.3	1.4
5.5	1.9
6.7	2.1
7.6	3.7
10.9	3.8
11.5	7.6
11.7	8.4
13.7	9.1
26.2	9.9

Correlation = 0.5

See next slide

Obtaining Sample Numbers to Create a Specified Correlation Structure

- Let $\mathbf{y}_1,\ldots,\mathbf{y}_n$ be the data sets, each length t
- Let **A** be the Cholesky factorization of correlation matrix **C**, where $\mathbf{A'A} = \mathbf{C}$
- Let $\mathbf{U}_{txn} = (\mathbf{u}_1,\ldots,\mathbf{u}_n)$, $\mathbf{u}_i$ {t uniform random num}
- Let $\mathbf{X} = \mathbf{U} \times \mathbf{A}$, Note $Var(\mathbf{Au}_i) = \mathbf{C}$
- Then $\mathbf{K}_{[,j]} = Rank(\mathbf{X}_{[,j]})$, $j = 1,\ldots,n$ are the sample numbers needed to generate the correlated aggregate result.

Additivity

- Means can be added
- Fractiles can be added ONLY when there is fractile additivity between distributions
- That's it

Provinces & Assessment Units

33 Provinces defined

69 Assessment Units evaluated

Quantitative estimates for 48 AUs

Data from 10,000 Monte Carlo Simulations

Trial	Risked Gas in Gas Fields (BCFG)	Risked Oil in Oil Fields (MMBO)
1	11,567	389
2	6,752	1,487
3	0	0
4	11,669	1,071
5	10,976	678

Circum Arctic Dependency Approach

- Ask assessors to specify pairwise correlations

- Assessment units close together tend to be more highly correlated than ones further distant

Problems!

- Given 48 assessment units, there are 48 x 47/2 = 1128 possible correlations
- Specifying pairwise correlations does not guarantee that the resulting matrix will be positive semi-definite

Circum Arctic Matrix

AU Code	AU Name	00010101	00010202	00020101	00020201	10080102	10080103	10500101	10500102	10500103
00010101	Makarov Basin Margin	1.00								
00010202	Siberian Passive Margin	0.70	1.00							
00020101	Lena Prodelta	0.20	0.27	1.00						
00020201	Nansen Basin Margin	0.20	0.20	0.30	1.00					
10080102	Main Basin Platform	0.20	0.20	0.20	0.20	1.00				
10080103	Foredeep Basins	0.20	0.20	0.20	0.20	0.80	1.00			
10500101	Kolguyev Terrace	0.20	0.20	0.20	0.20	0.80	0.80	1.00		
10500102	South Barents Basin and Ludlov Saddle	0.20	0.20	0.20	0.20	0.60	0.60	0.90	1.00	
10500103	North Barents Basin	0.20	0.20	0.20	0.20	0.50	0.50	0.80	0.90	1.00

Minimum eigenvalue = - 0.5

Solutions

- Adjust correlations unit matrix is positive semi-definite

- Specify distributions (beta or triangular) for pairwise correlation Frigessi A., and others, Quantitative Finance 11(7):1081-1090

- Use Bayes approach to guarantee that resulting matrix is positive semi-definite as it is specified

Minimize Frobenius Norm

- Projection system; Higham (2002, J. of Numerical Analysis)

$$\|\mathbf{A}-\mathbf{B}\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \left| a_{ij} - b_{ij} \right|^2 \right)^{1/2}$$

A is an improper matrix
B is the nearest correlation matrix

- R function nearcor in library sfsmisc
  - www.r-project.org

Small Example

Matrix of Correlations		
Var1	Var2	Var3
1	0.9	-0.3
0.9	1	0.3
-0.3	0.3	1

Eigenvalues		
Var1	Var2	Var3
1.900	1.168	-0.068

Correlation Matrix		
Var1	Var2	Var3
1	0.851	-0.273
0.851	1	0.273
-0.273	0.273	1

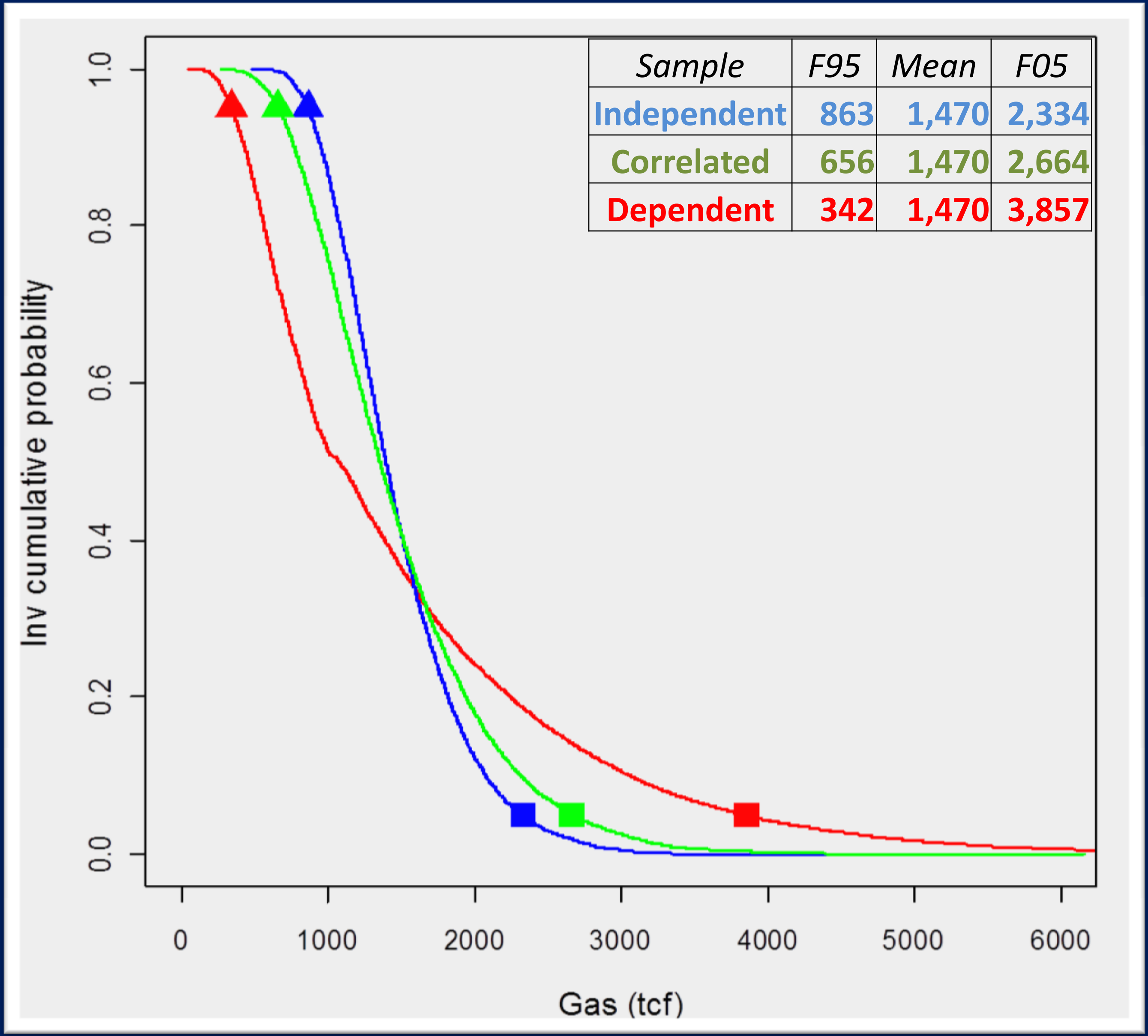
Frobenius norm = 0.0879

Max abs difference = 0.0488

Kaufman, Faith, and Schuenemeyer

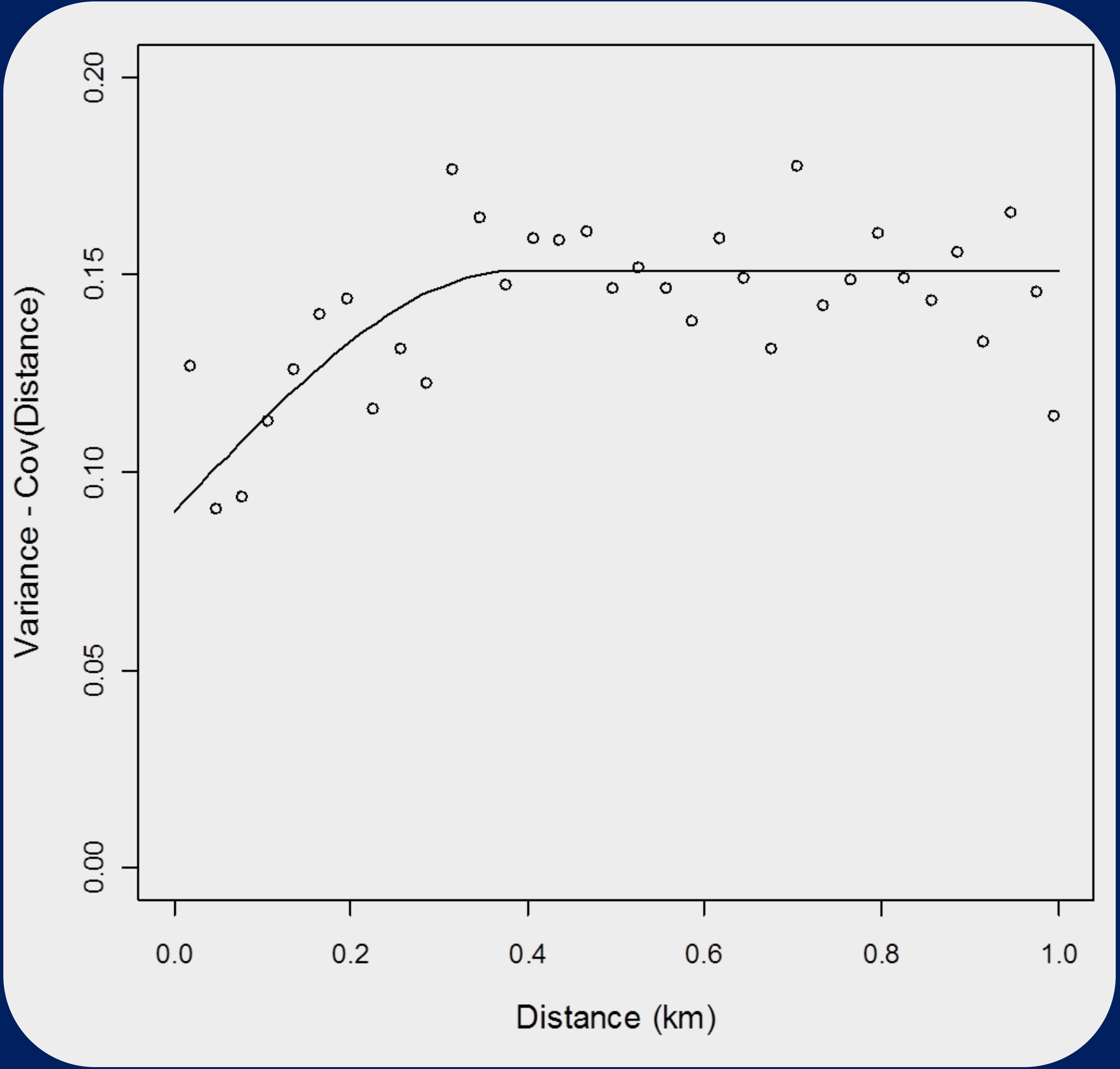
- In practice we ask geologists to assess
  - marginal distributions of magnitudes of hydrate accumulations in each unit under study
  - marginal distributions of the number of accumulations in each unit and
  - probabilistic dependencies among accumulations within and between units.

Dependency Matters!



Sample	*F95*	*Mean*	*F05*
Independent	**863**	**1,470**	**2,334**
Correlated	**656**	**1,470**	**2,664**
Dependent	**342**	**1,470**	**3,857**

Covariance (Semivariogram) Model



II. Data

1. Hard
2. Analogs

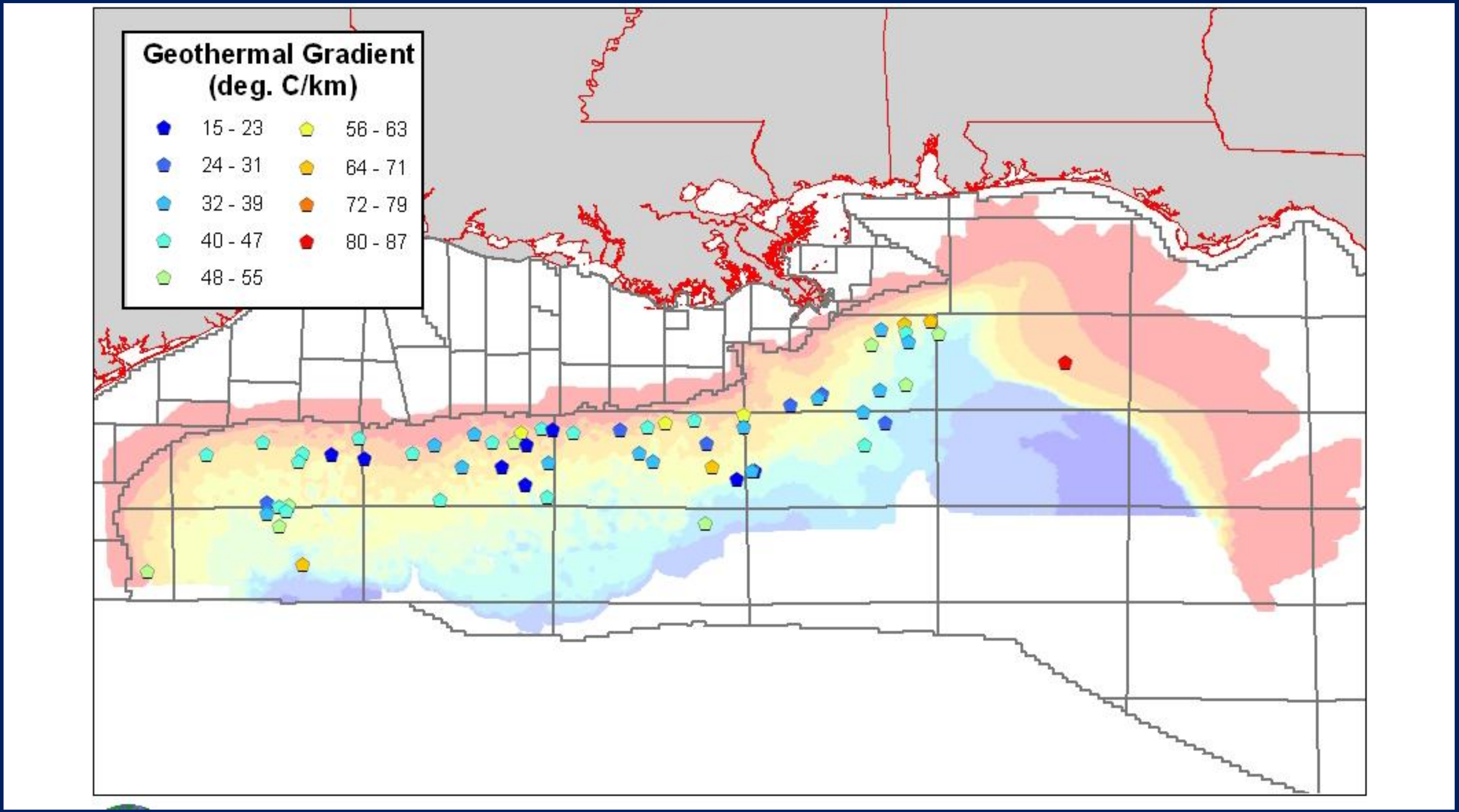
Why Analogs?

- Data expensive

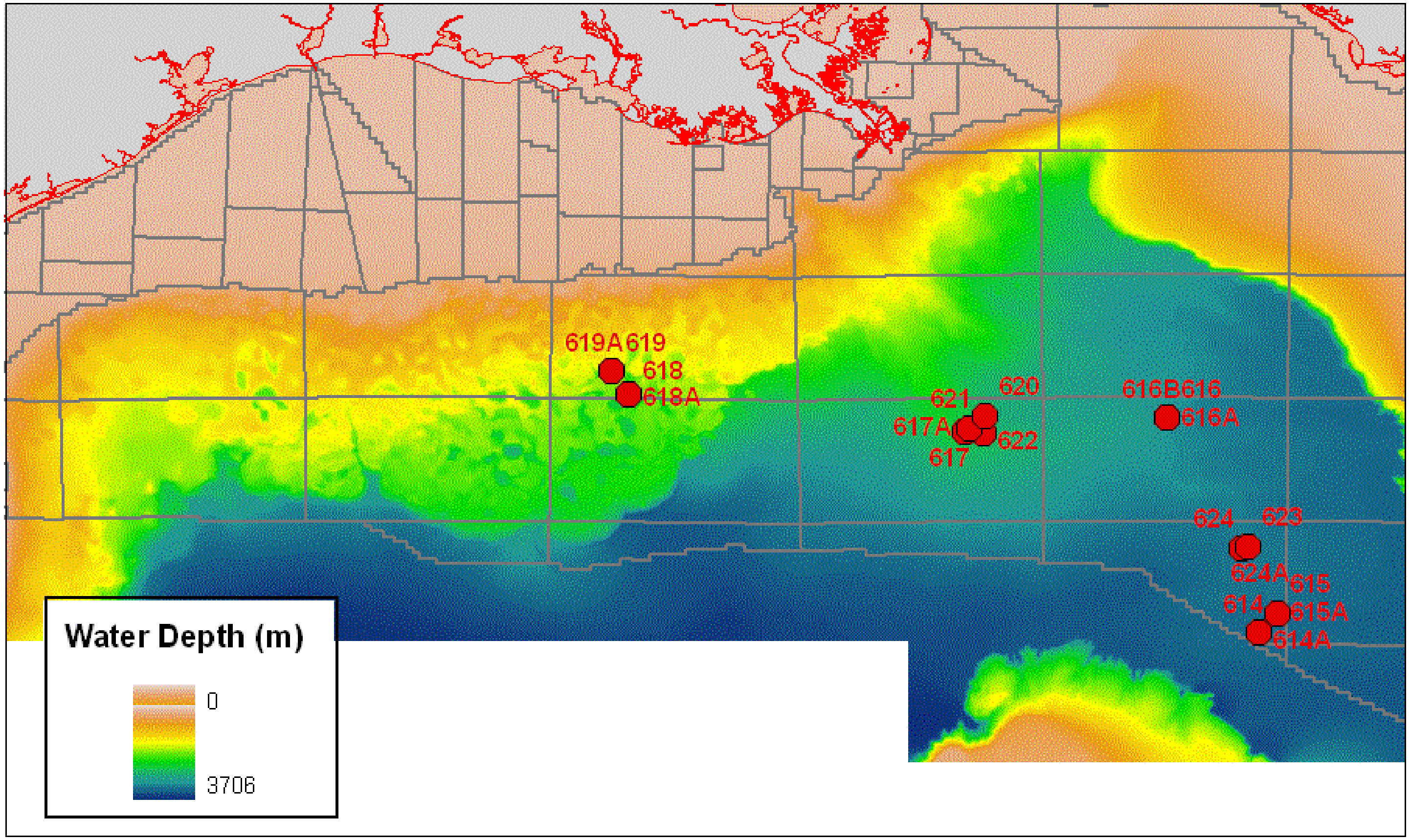
- In classical designed experiments, sample over area of interest and replicate – even here sometime analogs needed to estimate variance

- Examples:

Gulf of Mexico, Geothermal Gradient

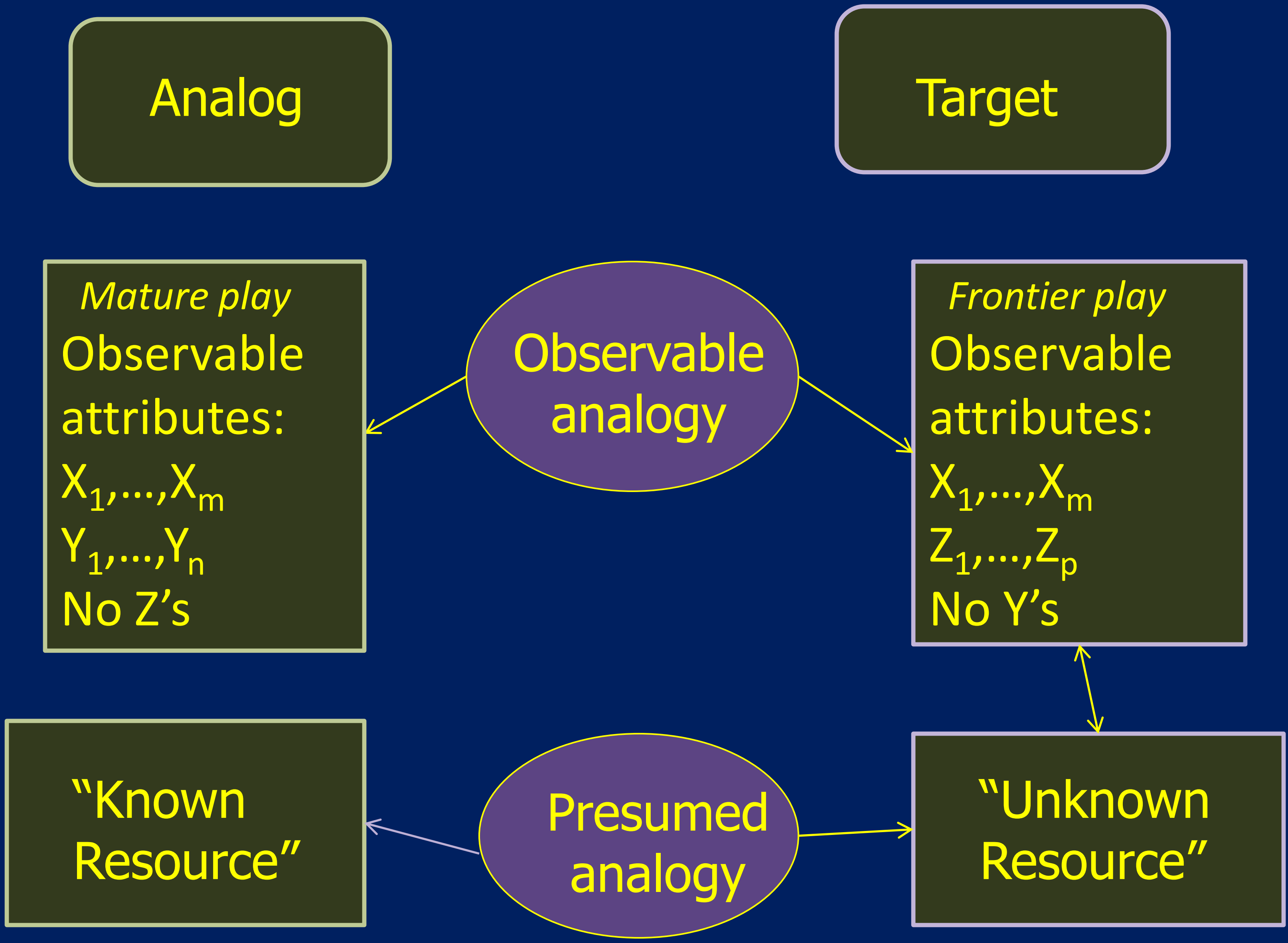


Total Organic Carbon Sites



Oil & Gas Resource Estimation

- Policy makers, energy companies, scientists, public
- Estimation of undiscovered resources
  - Level of exploration
    - Frontier to mature
  - Methods
    - Mature – Geological/statistical models – hard data
    - Frontier – Analogy, expert judgment


Analog
Target
Mature play
Observable attributes:
X1,…,Xm
Y1,…,Yn
No Z's
Observable analogy
Frontier play
Observable attributes:
X1,…,Xm
Z1,…,Zp
No Y's
"Known Resource"
Presumed analogy
"Unknown Resource"

US Geological Survey World Analog Database

- 246 Assessment Units (AUs)
- Observable attributes (factors) are nominal variables assigned to each AU:
  - Architecture
  - Trap system
  - Etc.
- Resources grouped by AU:
  - Sizes, numbers, & properties of oil and gas fields outside the U.S.
  - Includes ~ 95% of known petroleum (HIS, 2008 data)
  - Probabilistic estimates resources (USGS, 2000)

Observed Factors

Observable Attribute (Factor)	Max Number of Ordered Levels
Architecture	3
Trap System (Major)	4
Depositional System	4
Source Rock Depositional Environment	2
Kerogen Type	2
Source Type	2
Source Rock Qualifier	1
Status	1
Specific Reservoir Rock Age	1
General Reservoir Rock Age	1
Reservoir Rock Lithology	1
Reservoir Rock Depositional Environment	1
Seal Rock Lithology	1
Trap Type	1

Procedure Outline

- Identify observable attributes (factors) for inclusion via expert judgment (Don Gautier, USGS). In this example they are:
  – Architecture
  – Trap systems
  – Depositional systems
- Establish weighting scheme
  – All attributes are weighted equally
  – Levels within attributes are assigned decreasing weights

Architecture Levels (Arch_1, Arch_2 & Arch_3)

AU	Arch_1	Arch_2	Arch_3
38220101	Backarc	Strike-slip systems	Foreland
38220102	Backarc	Strike-slip systems	
38240101	Backarc	Strike-slip systems	
38240201	Backarc	Strike-slip systems	
38280101	Backarc	Strike-slip systems	
39100101	Rifted passive margin		
39100201	Rifted passive margin		

A Weighting Scheme

		Weights			
Factor	**Num of Levels**	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Architecture	**3**	5.333	3.333	1.333	
Trap System (Major)	**4**	4	3	2	1
Depositional System	**4**	4	3	2	1

Sampling Scheme

- 24 random samples (AUs) are selected without replacement from the analog database; two large AUs (by BOE) are added.
- Evaluation with a procedure is as follows:
  - Each of the 26 samples is, in turn, assumed to be the target AU
  - The remaining 122 AUs are candidate analogs to be compared with the target. Only AUs with > 50% resources estimated to have been discovered are considered.

Examples

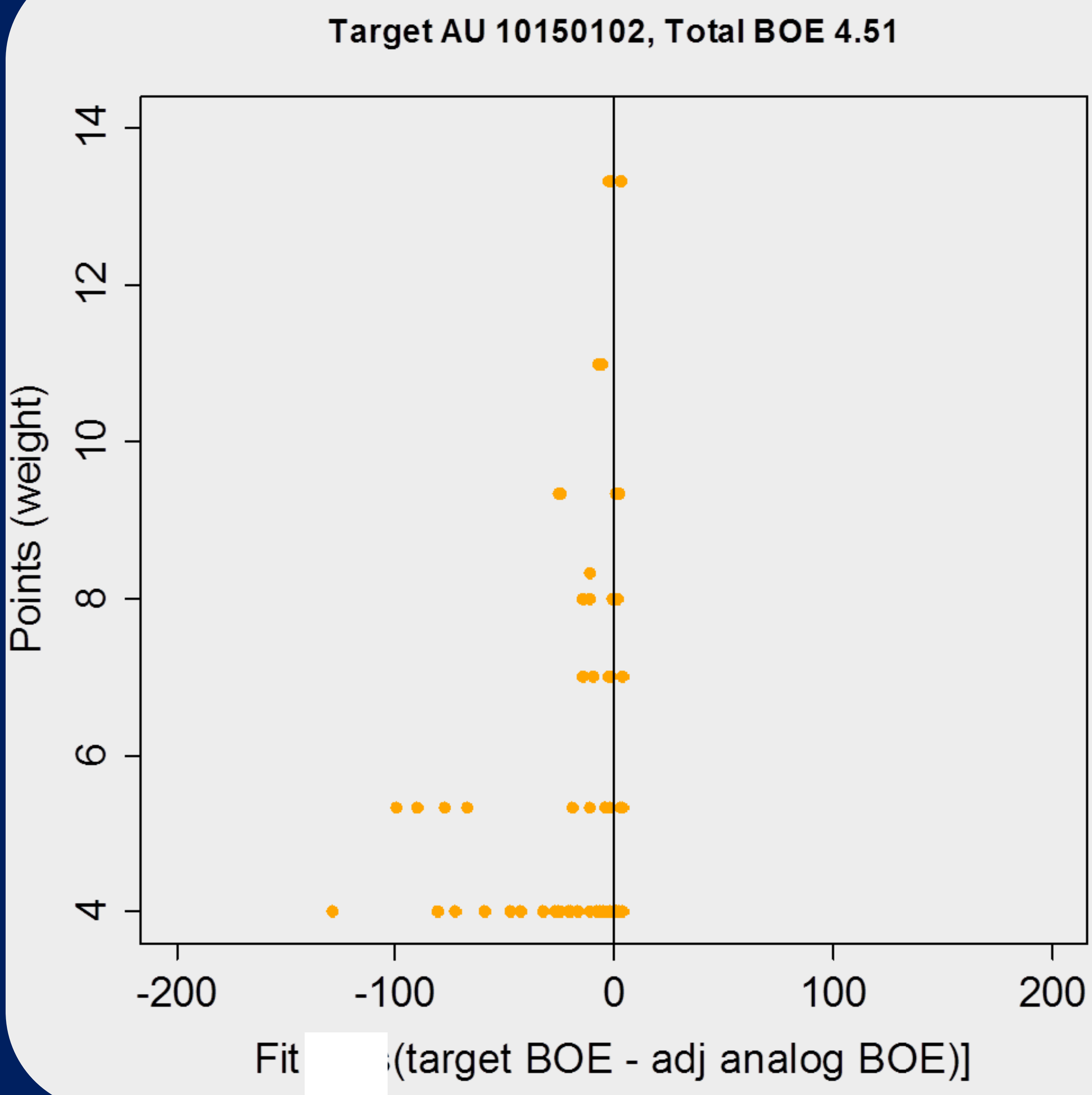
- Total BOE in analogs is rescaled to the area of the target (BOE density)

BOE Analog Resource Density =

Total Analog BOE x Target Area / Analog Area

Measures of Fit

- Many measures of fit
  - We use:
    - Target total BOE – density adjusted analog total BOE
  - Note we are assuming total BOE *known* in our model testing procedure


Target AU 10150102, Total BOE 4.51
Points (weight)
4
6
8
10
12
14
-200
-100
0
100
200
Fit
(target BOE - adj analog BOE)]

Analog Issues

- Using available information
- Missing data
- Uncertain data
- Structure of database
- Expert judgment
- Propagation of uncertainty

Using Available Information

- Attributes (factors)

Observable Attribute (Factor)	Max Number of Ordered Levels
Architecture	3
Trap System (Major)	4
Depositional System	4
Source Rock Depositional Environment	2
Kerogen Type	2
Source Type	2
Source Rock Qualifier	1
Status	1
Specific Reservoir Rock Age	1
General Reservoir Rock Age	1
Reservoir Rock Lithology	1
Reservoir Rock Depositional Environment	1
Seal Rock Lithology	1
Trap Type	1

Using Available Information

- Target resource may not be completely unknown. When possible use:
  - Known prospects or discoveries
  - Size-frequency distribution
  - Oil versus gas
- Multiple analogs
  - Can/should they be combined to provide more accurate results?

Missing & Uncertain Data

- Missing data: Discriminatory data elements may be missing
- Uncertainty: Example –
  - NE Greenland; no info; Broad regional characteristics; rift-sag basin covered entire area of assessment unit; density of resources in North Sea

Missing & Uncertain Data (continued)

- Suggestion: Designate via expert judgment, uncertainty in data elements
  - Analog resource base
    - Not all fields discovered
    - Estimates of remaining undiscovered
  - Areas (analog & target)
    - Uncertainty exists

The Database

- Biased (systematically wrong-how hierarchic assembled/assembled)

Propagation of Uncertainty

- Recall goal is to estimate undiscovered resources in target *and* provide an appropriate uncertainty estimate

- Uncertainty needs to reflect
  - Uncertainty in choice of analog database
  - Uncertainty in elements in database
  - Uncertainty associated with goodness of fit

II. 3. Expert Judgment

Examples

- Many disciplines use experts
  - Medicine
  - Food tasting
  - Economics
  - Geology – resource assessments
  - Climate
  - Hazards

Eliciting Expert Opinion

- Consensus
- Delphi
- Cooke, RM - calibration

Why Experts?

- Estimate future event
- Estimate event in present – measurement not feasible
  - Time
  - Money
  - Accessibility

What Do They Do?

- Answer questions like:
  - How long?
  - How much?

Concerns About Experts!

- Not all equal!
- Overconfident
- Calibrate or adjust for bias?
- In earth sciences – limited number
  - Different disciplines
- Weighting
- Gaming the system

Weighting

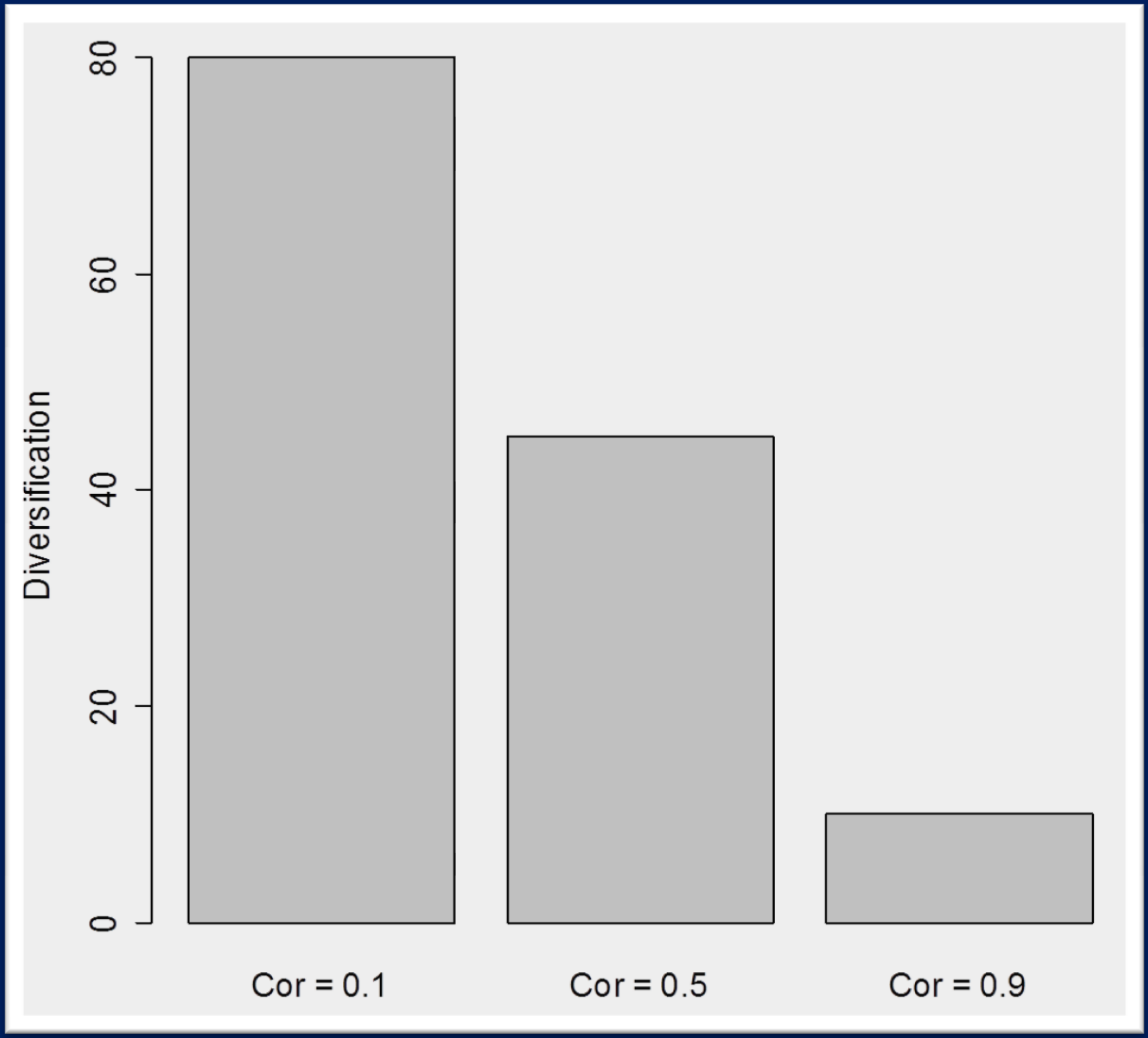
- Statistical
- Equal
- Self-selection

Thank you

- Questions – comments – suggestions

- Jack's contact info: jackswsc@q.com
- Southwest Statistical Consulting LLC: www.swstatconsult.com
- Statistics for Earth and Environmental Scientists: www.earthstatbook.com

Portfolio Management - Holdings



Consider a Mutual Fund

- Two types – balanced and specialized
- Specialized – energy, technology, health care
- Balanced fund – minimal correlation among holding
- Specialized fund – high correlation
- *It is essential to know the degree of dependency*

Uncertainty Intervals

- Producer perspective
  - W i d e
- Policy wonk
  - Narrow
- Investor
  - Realistic

Financial Risk - Reward

- Suppose investors need a 5% chance of at least 2,664 tcf gas
  - The “Correlated” scenario
  - Pr(Gas >= 2,664) = 0.05
- Alternative A: “Dependent” true
  - Pr(Gas >= 2,664) = 0.14
  - Okay but maybe not best use of resources
- Alternative B: “Independent” true
  - Pr(Gas >= 2,664) = 0.02
  - Could take significant loss

	Target	Analog (13.3 points)
AU_Code	Lower Volga	Western Pre-Aptian Reservoirs
Arch_1	Rifted passive margin	Rifted passive margin
Arch_2		
Arch_3		
TrapSys_1	Basement-involved block structures	Basement-involved block structures
TrapSys_2		
TrapSys_3		
TrapSys_4		
DepSys_1	Paralic clastics	Paralic clastics
DepSys_2	Carbonate shelf	Continental clastics
DepSys_3		
DepSys_4		
Area_sqkm	95,001	13,393
DiscBOE	4.248	0.767
UnDiscBOE	0.262	0.063
Tot Est Rec	4.51	0.83
Adj Est Rec	4.51	5.89

Trap Systems Levels

AU	TrapSys_1	TrapSys_2	TrapSys_3	TrapSys_4
80420102	Gravity-induced growth faults	Stratigraphic undeformed	Paleogeomorphic	
80430101	Basement-involved block structures	Stratigraphic undeformed		
80430102	Extensional grabens and other structures related to normal faulting	Stratigraphic undeformed		
80470201	Extensional grabens and other structures related to normal faulting	Basement-involved block structures	Stratigraphic undeformed	Gravity-induced growth faults
80470301	Stratigraphic undeformed	Gravity-induced growth faults		
80470302	Compressional anticlines, folds, thrusts	Gravity-induced growth faults		

Architecture Levels (Arch_1, Arch_2 & Arch_3)

AU	Arch_1	Arch_2	Arch_3
38220101	Backarc	Strike-slip systems	Foreland
38220102	Backarc	Strike-slip systems	
38240101	Backarc	Strike-slip systems	
38240201	Backarc	Strike-slip systems	
38280101	Backarc	Strike-slip systems	
39100101	Rifted passive margin		
39100201	Rifted passive margin		

A Weighting Scheme

		Weights			
Factor	Num of Levels	Level 1	Level 2	Level 3	Level 4
Architecture	**3**	5.333	3.333	1.333	
Trap System (Major)	**4**	4	3	2	1
Depositional System	**4**	4	3	2	1

Sampling Scheme

- 24 random samples (AUs) are selected without replacement from the analog database; two large AUs (by BOE) are added.

- Evaluation with a procedure is as follows:
  - Each of the 26 samples is, in turn, assumed to be the target AU
  - The remaining 122 AUs are candidate analogs to be compared with the target. Only AUs with > 50% resources estimated to have been discovered are considered.

Examples

- Total BOE in analogs is rescaled to the area of the target (BOE density)

BOE Analog Resource Density =

Total Analog BOE x Target Area / Analog Area

Examples –
3 in some detail; 23 quickly

- Example 1
- Target AU: Lower Volga

Area (sq km)	Total Est Recov BOE	Fraction discovered
95,001	4.51	0.94

Measures of Fit

- Many measures of fit
  - We use:
    - Target total BOE – density adjusted analog total BOE
  - Note we are assuming total BOE *known* in our model testing procedure

Examples –
3 in some detail; 23 quickly

- Example 1
- Target AU: Lower Volga

Area (sq km)	Total Est Recov BOE	Fraction discovered
95,001	4.51	0.94