USGS
science for a changing world

Global Mineral Resource Assessment

Aggregation of Estimated Numbers of Undiscovered Mineral Deposits—an R-Script with an Example from the Chu Sarysu Basin, Kazakhstan

Scientific Investigations Report 2010–5090–B

U.S. Department of the Interior
U.S. Geological Survey

This page left intentionally blank.

Global Mineral Resource Assessment

Michael L. Zientek and Jane M. Hammarstrom, editors

Aggregation of Estimated Numbers of Undiscovered Mineral Deposits—An R-Script with an Example from the Chu Sarysu Basin, Kazakhstan

By John H. Schuenemeyer, Michael L. Zientek, and Stephen E. Box

Scientific Investigations Report 2010–5090–B

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 2011

Suggested citation:
Schuenemeyer, J.H., Zientek, M.L, and Box, S.E., 2011, Global Mineral Resource Assessment—Aggregation of estimated numbers of undiscovered deposits—an R-script with an example from the Chu Sarysu Basin, Kazakhtan: U.S. Geological Survey Scientific Investigations Report 2010-5090-B, 13 p.

Contents

Figures

Tables

This page left intentionally blank.

Aggregation of Estimated Numbers of Undiscovered Mineral Deposits—An R-Script with an Example from the Chu Sarysu Basin, Kazakhstan

By John H. Schuenemeyer[1], Michael L. Zientek[2], and Stephen E. Box[2]

Introduction

Mineral resource assessments completed by the U.S. Geological Survey (USGS) during the past three decades express geologically based estimates of numbers of undiscovered mineral deposits as probability distributions. Numbers of undiscovered deposits of a given type are estimated in geologically defined regions. Using Monte Carlo simulations, these undiscovered deposit estimates are combined with tonnage and grade models to derive a probability distribution describing amounts of commodities and rock that could be present in undiscovered deposits within a study area. In some situations, it is desirable to aggregate the assessment results from several study areas (Schuenemeyer, 2003, 2005). This report provides a script developed in open-source statistical software, R (R Development Core Team, 2010), that aggregates undiscovered deposit estimates of a given type, assuming independence, total dependence, or some degree of correlation among aggregated areas, given a user-specified correlation matrix.

Assessment Method

The USGS uses a three-part form of assessment to estimate numbers of undiscovered mineral deposits within a geologically defined study area. This form of assessment is based on mineral deposit models. Descriptive mineral deposit models document the geologic criteria for delineating permissive geographic areas (permissive tracts) for the occurrences of deposits of specific types. Grade and tonnage models have the form of frequency distributions of average grade and size of thoroughly explored examples of each type of mineral deposit. For most models, the distributions are positively skewed and can be approximated by a lognormal distribution.

The probable amounts of mineral resources associated with undiscovered deposits by deposit type may be estimated by combining a probability distribution for the number of undiscovered deposits of a given type within a tract with appropriate grade and tonnage distributions in a Monte Carlo simulation (Root and others, 1992; Singer, 1993; Singer and Menzie, 2010).

The distribution of undiscovered deposits is estimated by expert panels and used to create a cumulative distribution of discrete values. For most assessments, the number of undiscovered deposits is elicited at 3 to 5 probability percentiles (90th, 50th, 10th, 5th, and 1st). For example, an estimate at the 90th percentile is the greatest number of deposits present with a probability 0.9 or more for a given deposit type in a permissive tract; that is, the probability of at least that many deposits being present is 0.9 or greater, and the probability of more deposits is less than 0.9. The estimate made at the 90th percentile is the number of undiscovered deposits for which panel members are the most confident. The estimated number of undiscovered or more deposits at the 50th percentile has an even chance of being present (p=0.5). Estimates at lower percentiles can be considered long shots. However, estimated numbers of deposits, even at very low probabilities (for example 1 chance out of 100) need to be reasonable given the size of the tract and possible indications of mineralization suggested by prospects, occurrences, and other data. The estimators try to estimate nonzero values for at least three percentiles (for example, 90th-50th-10th is 1-1-1, or 90th-50th-10th-5th-1st is 0-0-1-1-1).

In 2010, an expert panel assessed undiscovered resources associated with sandstone copper deposits that may occur in the Chu Sarysu Basin, Kazakhstan. The geology and genesis of sandstone-copper-type deposits are reviewed by Cox and others (2003) and Hitzman and others (2010). In this basin, sandstone copper deposits are localized on anticlinal structures, where hydrocarbons may have accumulated prior to the migration of copper-enriched, oxidized brines. The assessment panel evaluated structural traps along with physical evidence for the migration and accumulation of hydrocarbons and later copper-enriched, oxidized brines. Six tracts were delineated, corresponding to 6 undiscovered deposit estimates for 6 prominent structural features in the Tesbulak subbasin, and a seventh estimate for the remaining part of the subbasin lacking these features (fig. 1; table 1).

[1]Southwest Statistical Consulting, Cortez, Colorado.

[2]U.S. Geological Survey, Spokane, Washington.

Table 1. Estimated number of undiscovered sandstone copper deposits in the Tesbulak subbasin, Chu Sarysu Basin, Kazakhstan, at 5 percentiles.

Tract name	Tract	Percentiles 90	50	10	5	1
Zhaman-Aibat	T1	3	4	5	5	5
Central Sarysu uplift	T2	0	1	3	5	5
Kulen (ZA1)	T3	0	1	2	2	3
Zhaktyktau (ZA-3)	T4	0	0	1	2	2
East Karakoin (ZA-4)	T5	0	1	1	2	5
Dautbay	T6	0	1	2	3	5
Chu-Sarysu north, beyond	T7	0	2	2	5	5

Once estimates are made by the panel, a cumulative discrete probability distribution for undiscovered deposits is calculated. An infinite number of distributions are consistent with the 3 to 5 estimates of undiscovered deposits made at various percentile values, as estimated by the expert panel. For Monte Carlo simulation, a default distribution is chosen that is approximately in the middle of all possible choices (Root and others, 1992). As the program is currently configured (Root and others, 1992), the largest number of estimated undiscovered deposits at the lowest probability also is the maximum number of deposits predicted in the simulation. The default distribution derived from the data in table 1 is shown in table 2; the allocation of the unit probability among the nonnegative integers that define the default distribution of the number of deposits is described by Root and others (1992) and is part of the output from the simulation software (Root and others, 1996; Duval, 2004; Baweic and Spanski, in press). This probability distribution of undiscovered deposits is combined with the grade and tonnage models using Monte Carlo simulations to calculate amounts of undiscovered metals, materials, and mineralized rock (Root and others, 1992).

Table 2. Default probability distribution of number of deposits (n) as determined by Mark3/EMINERS simulator for the estimates given in table 1.

n	T1	T2	T3	T4	T5	T6	T7
0	0.0286	0.3	0.3	0.7	0.3	0.3	0.2
1	0.0286	0.3	0.4	0.225	0.625	0.4	0.2
2	0.0286	0.2	0.27	0.075	0.0317	0.225	0.5083
3	0.2142	0.1125	0.03		0.0133	0.035	0.0167
4	0.4	0.025			0.0133	0.02	0.0167
5	0.3	0.0625			0.0167	0.02	0.0583

Overall, how many undiscovered deposits are predicted given the estimated numbers of undiscovered deposits in the 6 subunits and for the balance of the subbasin? For the assessment report, the expert panel decided to report undiscovered resources by subbasin. To arrive at this single set of estimates for the entire Tesbulak subbasin, a process to statistically aggregate the 7 assessment estimates in table 2 is required. This combined estimate of numbers of undiscovered deposits can, in turn, be combined with grade and tonnage distributions to prepare predicted distributions of undiscovered copper and rock.

Statistics of Aggregation

The statistical approach to aggregation used in this report is discussed by Schuenemeyer and Drew (2011) and is briefly summarized here. Consider a simple example where X and Y are random variables representing the distributions of undiscovered deposits in two assessment units characterized by probability functions f_x and f_y:

$$X \sim f_x(\mu_X, \sigma_X^2) \text{ and } Y \sim f_y(\mu_Y, \sigma_Y^2)$$

where μ_X and μ_Y are the means and σ_X and σ_Y are the standard deviations. The distribution of the sum of these two functions is given by:

$$X+Y = f_{x+y}(\mu_{X+Y}, \sigma_X^2 + \sigma_Y^2 + 2Cov(X,Y))$$

where $Cov \neq 0$ is the correlated variation of X and Y. Dependencies between distributions do not affect the mean of the distributions, only the spread (as indicated by the standard deviation). The mean of the aggregated distributions is the sum of the means of the individual distributions:

$$\mu_{X+Y} = \mu_X + \mu_Y$$

However, aggregation does affect the spread of the functions because the variance of the combined distribution is affected by the dependency between the random variables:

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2Cov(X,Y)$$

Independence implies that the occurrence of one event makes it neither more nor less probable that the other event occurs. Under this condition of independence,

$$Cov(X,Y) = 0,$$

and the variance of the combined distributions is:

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2.$$

For dependence, events in one region predict the events in a second region, such as number of deposits. Thus,

$$Cov(X,Y) = \rho_{XY}\sigma_X\sigma_Y,$$

where ρ_{XY} is the correlation coefficient between X and Y. If $\rho_{XY} = 1$, then the two distributions are totally dependent, which implies that percentiles can be added.

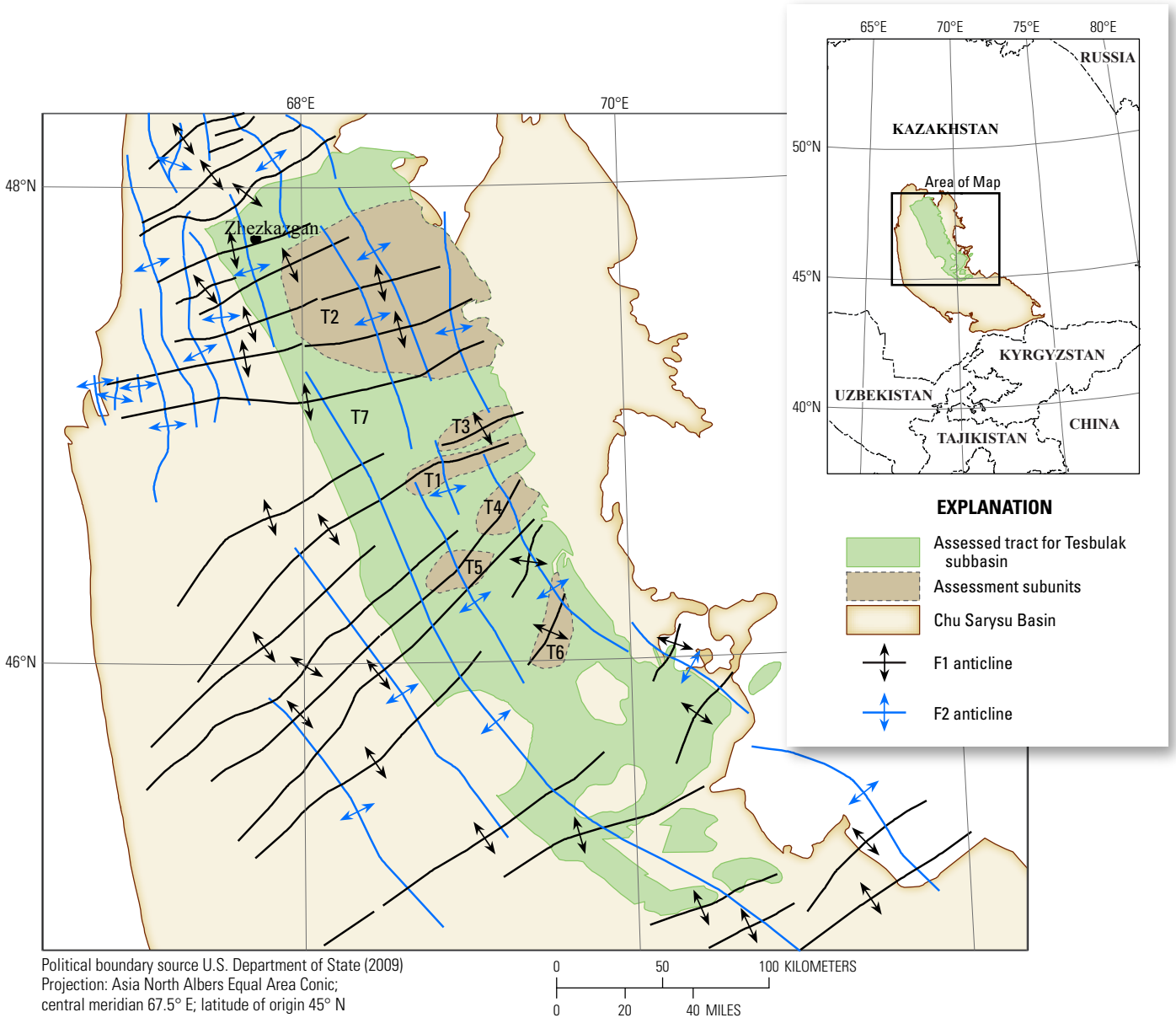


Figure 1. Map showing the location of the assessment areas (assessment subunits) with undiscovered deposit estimates within the Tesbulak subbasin (including the balance of the subbasin as assessment tract T7), Chu Sarysu Basin, Kazakhstan.

Generally, dependence (association) is positive, but negative association also is permissible:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho_{XY}\sigma_X\sigma_Y.$$

Thus, the variance of $X+Y$ is a maximum for total positive dependence. For the Tesbulak subbasin, geologists on the expert panel decided that some, but not total, dependence between predicted deposits is likely in different tracts. An aggregation code presented in this report combines undiscovered deposit estimates and reports the aggregated total, assuming independence, total dependence, and positive dependence based on user-defined subjective pairwise correlations between each assessed area. The approach is similar to aggregation methodology described by Schuenemeyer (2003, 2005).

Aggregation Code to Combine Undiscovered Deposit Estimates

The aggregation codes requires two input files: (1) a file consisting of tract identifier, number of deposits, and probability (table 3); and (2) a file consisting of a correlation matrix (table 4). Both must be comma-separated values (csv) files and must be saved in a folder (in the example called AggT), which also contains the R document that contain the scripts. The input files are described below and are listed in appendixes 1 and 2. The aggregation functions are given in appendixes 3 and 4. The process to install R and run the script is given in appendixes 5 and 6.

Input File 1—Number of Deposits and Associated Probabilities

Data for the first csv input file consists of a tract ID (TID), number of deposits (n), and the corresponding probability (Pr). These data are provided in the EMINERS output that displays during EMINERS execution (Duval, 2004; Bawiec and Spanksi, in press), or in other software that reproduces the distribution, as defined by Root and others (1992). If using EMINERS, the raw output may be saved by "selecting all" and copying the display to an Excel worksheet that can be edited to create input file 1. The first two tracts of a sample file based on table 2 are shown in table 3. The entire file is given in appendix 1. The sum of probabilities within a tract must add to one. For this example, the file is called newCS2.csv (the name is arbitrary, but it must be a csv file).

Table 3. Input file for tract ID (TID), number of deposits (n), and corresponding probability (Pr).

TID	n	Pr
T-1	0	0.0286
T-1	1	0.0286
T-1	2	0.0286
T-1	3	0.2142
T-1	4	0.4
T-1	5	0.3
T-2	0	0.3
T-2	1	0.3
T-2	2	0.2
T-2	3	0.1125
T-2	4	0.025
T-2	5	0.0625

Input File 2—Correlation Matrix of Assessor-Defined Dependencies

This input file is a "correlation matrix", as shown in table 4 and appendix 2. There are seven tracts in the example. Pairwise correlations were specified by assessors for seven tracts to be aggregated based on geologic understanding of the deposit model of sandstone copper deposits (Cox and others, 2003; Hitzman and others, 2010). Measures of dependencies of the predicted deposits among tracts are related to sources of copper, basin fluids, reductants, and structural traps. If assessment areas (plays) are on the same anticline, along the same up dip stratigraphic pinchout trap, along the same fault trap, and so on, the higher the dependence between the assessment areas. For example, the two assessment subunits on the same structural feature (T4 and T5 in fig. 1) would be considered to have a stronger degree of dependence; given the presence of a deposit in one subunit, the probability for a deposit in the second subunit would increase. As a general rule, the expectation is that nearby assessment tracts will tend to be more highly correlated than those further apart.

Table 4. Correlation matrix used to aggregate undiscovered deposit estimates for tracts in tables 1 and 2.

	T1	T2	T3	T4	T5	T6	T7
T1	1						
T2	0.5	1					
T3	0.75	0.5	1				
T4	0.6	0.2	0.6	1			
T5	0.6	0.2	0.6	0.75	1		
T6	0.2	0.2	0.2	0.6	0.5	1	
T7	0.2	0.2	0.2	0.2	0.2	0.2	1

The upper diagonal does not need to be filled in. For this example, the matrix is called UsrCorr.csv (again the name is arbitrary, but it must be a csv file).

There is no guarantee that a matrix resulting from the specification of pairwise correlation will be a correlation matrix. To be a correlation matrix, UsrCorr.csv must be positive definite. This is equivalent to the smallest eigenvalue being positive. Positive definiteness assures that the pairwise correlations are logically consistent. For the matrix shown in table 4, the smallest eigenvalue is 0.18, thus, the matrix is consistent with a valid a correlation matrix. If the matrix is not positive definite, algorithm AggtEx.fn makes a bias adjustment, and the resulting correlation matrix is written to a file called BiasCorr.csv and put in a folder (AggT, or whatever it is named). This is accomplished by R function AggtEx.fn. The BiasCorr.csv correlation matrix can be compared with the original matrix UsrCorr.csv to determine the amount of bias.

Results

The output file generated by AggtEx.fn, called AgCS2Ex.csv, is shown in table 5. The first row (Indep) gives the results assuming independence; the third row (Total Dep) shows the results assuming total dependence. Results obtained using the user-specified correlation matrix (Correlation) are given in the second row. Note that the mean values for all three rows of table 5 are the same; however, the standard deviations and the coefficients of variation (CV) increase from independence to correlation to total dependence. These aggregated, undiscovered deposits estimates were combined with the grade and tonnage distribution for sandstone copper deposits by using Monte Carlo simulation. Summary statistics of the resulting contained metal distributions for copper are given in table 6.

Table 5. Output from the aggregation function.

Tracts	Assoc	P10	P50	P90	P95	P99	Mean	Std_Dev	CV
7	Indep	7	10	14	15	18	10.2805	2.792798	0.27166
7	Correlation	5	10	16	18	22	10.2805	4.250277	0.413431
7	Total Dep	3	10	16	24	30	10.2805	6.517059	0.633924

Table 6. Results of Monte Carlo simulations of undiscovered resources in the Tesbulak subbasin, Kazakhstan.

Material	Probability of at least the indicated amount					Mean	Probability of	
	0.95	0.9	0.5	0.1	0.05		Mean or greater	None
Independent								
Cu (T)	1,900,000	3,700,000	20,000,000	46,000,000	55,000,000	24,000,000	0.43	0.01
Correlated								
Cu (T)	1,300,000	2,900,000	20,000,000	50,000,000	61,000,000	24,000,000	0.43	0.02
Total dependence								
Cu (T)	540,000	1,600,000	19,000,000	53,000,000	66,000,000	24,000,000	0.41	0.03

Discussion

The amount of undiscovered metals and mineral materials often is of interest to those whose area of concern may be the subject of multiple mineral resource assessments. Therefore, there is a need to be able to aggregate the outcomes of multiple assessments into a single result. This can be done by statistically aggregating the numbers of undiscovered deposits predicted from all tracts. To do so requires assessors to understand the implications of assumptions made about the degree of association between various shared factors in geologically based assessment regions and tracts before aggregating assessment results. The key to aggregating results in the Chu Sarysu study is based on an understanding of the physical evidence for ore-forming processes as determined from known deposits in the area. Many of the assessment tracts share the same sources of copper and oxidized brines and are hosted in the same reservoir facies rocks. Some assessment units are on the same structural trap and likely shared the same reductant, which controls the deposition of copper minerals. For purposes of aggregation, an assumption of independence will yield estimates of aggregate uncertainty that are unrealistically small given the predicted deposits are not independent. Conversely, an assumption of total dependence will yield estimates of uncertainty that often are unrealistically large given the predicted deposits are dependent to some degree. The algorithm presented in this report allows user-specified correlation to be used to yield estimates of aggregated uncertainty where some degree of dependence is likely and can be estimated.

Acknowledgments

The example in this report is one outcome of a USGS assessment workshop on the Chu Sarysu Basin held in Vancouver, Washington, in October 2010. Members of the assessment team included: Steve Box (USGS), Boris Syusyura

(Mining and Economic Consulting, Ltd.), Vladimir Chechetkin (Zabaikalsky Division, Russian Geological Society), Reimar Seltman (Centre for Russian and Central EurAsian Mineral Studies, Natural History Museum, London), Murray Hitzman (Colorado School of Mines), Timothy Hayes (USGS), Cliff Taylor (USGS), and Michael Zientek (USGS). Pam Cossette and John Wallis, both with USGS, compiled the GIS database used for the assessment; Alla Dolgopolova (Natural History Museum, London) helped with translation.

The aggregation script was prepared for a workshop by Southwest Statistical Consulting, LLC, for the USGS in May 2010. The workshop was organized after discussions with Don Gauthier (USGS) on aggregation strategies used for USGS energy assessments.

Jane Hammarstrom (USGS) provided comments on an early version of this manuscript, and technical reviews were provided by Gilpin R. Robinson and James D. Bliss, both with the USGS.

References Cited

Adler, Joseph, 2010, R in a nutshell—A desktop quick reference: Sebastopol, Calif., O'Reilly Media, 640 p.

Bawiec, W.J., and Spanski, G.T., in press, EMINERS—Economic Mineral Resource Simulator, version 3.0: U.S. Geological Survey Open-File Report 2009-1057, program files and 29-p. Quick-Start Guide.

Cox, D.P., Lindsey, D.A., Singer, D.A., and Diggles, M.F., 2003 [revised 2007], Sediment-hosted copper deposits of the world—Deposit models and database: U.S. Geological Survey Open-File Report 2003-107 version 1.1, 50 p., accessed January 31, 2011, at http://pubs.usgs.gov/of/2003/of03-107/.

Duval, J.S., 2004, Version 2.0 of EMINERS—Economic Mineral Resource Simulator: U.S. Geological Survey Open-File Report 2004-1344 accessed January 31, 2011, http://pubs.usgs.gov/of/2004/1344/. [This report is temporarily unavailable. March 3, 2006].

Hitzman, M.W., Selley, David, and Bull, Stuart, 2010, Formation of sedimentary rock-hosted stratiform copper deposits through Earth history: Economic Geology, v. 105, no. 3, p. 627-639.

Murrell, Paul, 2006, R Graphics: Boca Raton, Flor., Chapman & Hall/CRC Computer, 328 p.

R Development Core Team, 2010, R—A language and environment for statistical computing: accessed February 5, 2011, at http://www.R-project.org.

Root, D.H., Menzie, W.D., and Scott, W.A., 1992, Computer Monte Carlo simulation in quantitative resource estimation: Nonrenewable Resources, v. 1, no. 2, p. 125-138.

Root, D.H., Scott Jr., W.A., and Selner, G.I., 1996, Computer program for aggregation of probabilistic assessments of mineral resources: U.S. Geological Survey Open-File Report 96-094.

Schuenemeyer, J.H., 2003, Methodology and results from the assessment of oil and gas resources, National Petroleum Reserve, Alaska: U.S. Geological Survey Open-File Report 03-118, 201 p., accessed January 31, 2011, at http://pubs.usgs.gov/of/2003/of03-118/.

Schuenemeyer, J.H., 2005, Methodology for the 2005 USGS assessment of undiscovered oil and gas resources, Central North Slope, Alaska: U.S. Geological Survey Open-File Report 2005-1410, 82 p. (http://pubs.usgs.gov/of/2005/1410/).

Schuenemeyer, J.H., and Drew, L.J., 2011, Statistics for earth and environmental scientists: Hoboken, New Jersey, John Wiley and Sons, Inc., 407 p.

Singer, D.A., 1993, Basic concepts in three-part quantitative assessments of undiscovered mineral resources: Nonrenewable Resources, v. 2, no. 2, p. 69-81.

Singer, D.A., and Menzie, W.D., 2010, Quantitative mineral resource assessments—An integrated approach: New York, Oxford University Press, 219 p.

Vernables, W.N., Smith, D.N., and the R Development Core Team, 2010, An introduction to R—Notes on R—A programming environment for data analysis and graphics, version 2.12.1 (2010-12-16): accessed January 28, 2011, at http://cran.r-project.org/doc/manuals/R-intro.pdf, 101 p.

Appendixes 1–6

Appendix 1. Input Tract File newCS2.csv

TID	n	Pr
T1	0	0.0286
T1	1	0.0286
T1	2	0.0286
T1	3	0.2142
T1	4	0.4
T1	5	0.3
T2	0	0.3
T2	1	0.3
T2	2	0.2
T2	3	0.1125
T2	4	0.025
T2	5	0.0625
T3	0	0.3
T3	1	0.4
T3	2	0.27
T3	3	0.03
T4	0	0.7
T4	1	0.225
T4	2	0.075
T5	0	0.3
T5	1	0.625
T5	2	0.0317
T5	3	0.0133
T5	4	0.0133
T5	5	0.0167
T6	0	0.3
T6	1	0.4
T6	2	0.225
T6	3	0.035
T6	4	0.02
T6	5	0.02
T7	0	0.2
T7	1	0.2
T7	2	0.5083
T7	3	0.0167
T7	4	0.0167
T7	5	0.0583

Appendix 2. User-Defined Correlation Matrix

	T1	T2	T3	T4	T5	T6	T7
T1	1						
T2	0.5	1					
T3	0.75	0.5	1				
T4	0.6	0.2	0.6	1			
T5	0.6	0.2	0.6	0.75	1		
T6	0.2	0.2	0.2	0.6	0.5	1	
T7	0.2	0.2	0.2	0.2	0.2	0.2	1

Appendix 3. Aggregation Function

```
AggtEx.fn<-function(Exfn,UsrC,nt=2000){
# Aggregation of user specified discrete distributions
# Exfn is user data (ID, n and Prob)
# UsrC is user specified correlation matrix
# nt is number of trials for simulation run

  uv<-unique(Exfn[,1])
  nau<-length(uv)
  ri<-matrix(0,nt,nau)
  squ<-matrix(0,3,10)
# generate distributions
  for (i in 1:nau){
# print(i)
  urn<-runif(nt)
  rv<-Exfn[Exfn[,1]==uv[i],]
  # cumulative distribution
  cus<-cumsum(rv[,3])
  lc<-length(cus)
    for (j in 1:nt){
     urn1<-urn[j]
     for (k in 1:lc){
       if(urn1 < cus[k]) {
        ri[j,i]<-rv[k,2]
        break}
    }
   }
  }
```

```
# independence
    t2<-rowSums(ri[,1:nau])
    t1<-Sum.fn(t2)
    squ[1,]<-c(nau,"Indep",t1)

#sort distributions
for (i in 1:nau){
   ri[,i]<-sort(ri[,i])
  }

or<-length(UsrC[,1])
orw<-length(UsrC[1,])
if(orw>or)  UsrC<-UsrC[,2:orw]
if(or != nau){
   print(c("num tracts",nau," not equal order corr matrix",or))
   stop}
for (i in 1:(or-1)){
  for (j in (i+1):or) {
  UsrC[i,j]<-UsrC[j,i]
}}

# uniform numbers for correlation
   rv<-runif(nau*nt,-1,1)
   U<-matrix(rv,nt,nau)
   or<-length(UsrC[,1])
   for (i in 1:(or-1)){
    for (j in (i+1):or) {
     UsrC[i,j]<-UsrC[j,i]
   }}

  #print(UsrC)
  t2<-as.matrix(UsrC)
  eig<-eigen(t2)
  eval<-eig$values
  #print(eval[nau])
 # is matrix a correlation matrix
  if(eval[nau] <= 0){
 # adjust matrix to be correlation
    bias<-abs(eval[nau])+0.001
    eval<-eval+bias
    evec<-eig$vectors
    t2<-evec%*%diag(eval)%*%t(evec)
    tri<-t2[1,1]
    t2<-t2/tri
     for(k in 2:nau){
       for(k1 in 1:(k-1)){
```

```
      t2[k,k1]<-t2[k1,k]
    }
    }
  #print(t2)
   write.csv(t2,file="BiasCorr.csv",row.names=FALSE)
   }

 Ch<-chol(t2)
 V<-U%*%Ch
 Ags<-0
  for (j in 1:nau){
  t3<-rank(V[,j])
  t4<-ri[,j]
  t5<-t4[t3]
  Ags<-Ags+t5
  }
 t1<-Sum.fn(Ags)
 squ[2,]<-c(nau,"Correlation",t1)

#totally dependent
 Ags<-rowSums(ri[,1:nau])
  t1<-Sum.fn(Ags)
  squ[3,]<-c(nau,"Total Dep",t1)
#output
 squ<-as.data.frame(squ)
 names(squ)<-c("Tracts","Assoc","P10", "P50", "P90", "P95", "P99",
"Mean", "Std_Dev", "CV")
 write.csv(squ,file="AgEx.csv",row.names=FALSE)
 print("Output is in AgEx.csv")
 }
```

Appendix 4. Sum Function

```
Sum.fn<-function(da){
 mri<-mean(da)
 sdri<-sd(da)
 cvri<-sdri/mri
 va<-c(quantile(da,c(0.10,0.50,0.90,0.95,0.99)),mri,sdri,cvri)
 return(va)
}
```

Appendix 5. R-Language and Software Installation

R is an open source code language that creates an environment for statistical computing and graphics. It contains numerous statistical techniques, including regression, spatial analysis, time-series analysis, clustering and classification. A variety of one-, two-, and three-dimensional graphical procedures is available. In addition, it can produce high-quality, publication-ready graphics. It has an easy-to-use language so that individuals can write their own programs.

The R language can read data from Excel csv and other file structures. It can interface C, C++, and Fortran code.

The purpose of this introduction is to have you gain a basic familiarity with R, as we will use it to illustrate aspects of aggregation. If you are not familiar with R, please review this document before attempting to implement the code.

Manuals

Numerous free manuals are available at the r-project Website (http://www.r-project.org/). The best one to begin with is ***An Introduction to R*** by Vernables and others (2010). Specific packages also have manuals. Numerous books have been published on the R-language, applications, and graphics (including Adler, 2010 and Murrell, 2006). A Website illustrating many graphics applications is http://addictedtor.free.fr/graphiques/.

Notes and a few simple commands:

- The > is the R prompt.
- The replacement command is the symbol <- which consists of two characters, the < followed by a -, for example x<- 3 + 4
- If you want to retrieve a previous line, hit the up arrow; if you went too far hit the down arrow on your keyboard.
- Generally, text goes between quotes. Note that these quotes are different than what you type in Word, for example, “text”.
- To see what is in your R directory type ls**( ).**
- To print a file in your R directory type the file name, for example, Ctest and hit enter.
- You cannot re-execute a series of commands in R by doing a copy-paste within R because you will be copying the >.
- Commands are not stored in R when you do a save worksheet, only functions (for example, CLT.fn) and files (for example, Ctest, CtestV, and Cbox) will be saved. We suggest that you save your R commands in a Notepad (or equivalent) file.
- To find the length of a vector, for example, CtestV, type length(CtestV).
- To obtain the dimensions of a matrix, for example, Ctest, type dim(Ctest).

Help

- When in R, click on Help.
- Click on R functions (text) if you know the name of the command, such as sample.
- Click on Search help if you know the general topic.
- Other help options are available such as FAQ and Manuals.

Installing R

To install R on your computer:

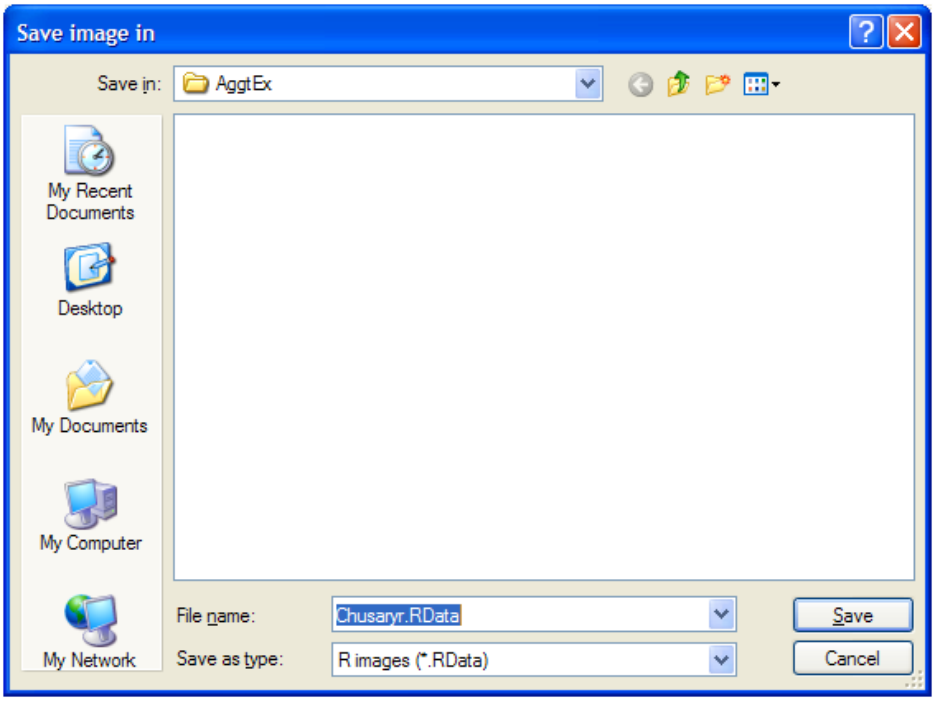
1. From the internet go to www.r-project.org.
2. Left click on on the link, CRAN (left side of screen under Download, Packages).
3. Scroll down to USA (or other country) and left click on a USA site.
4. In Downloads and Install R box you will see Linux, MacOS X and Windows. Left click on your operating system.
5. Windows users click on the link, base; MAC users click on appropriate version.
6. Follow the appropriate installation instructions. Note that the version as of January 31, 2011, is R 2.12.1. As updates occur regularly, a later version may be available when you install it. A nice feature of R is compatibility with previous versions.
7. Generally follow the defaults.

Appendix 6.
Executing the R-Code for Aggregation of Undiscovered Deposit Distributions

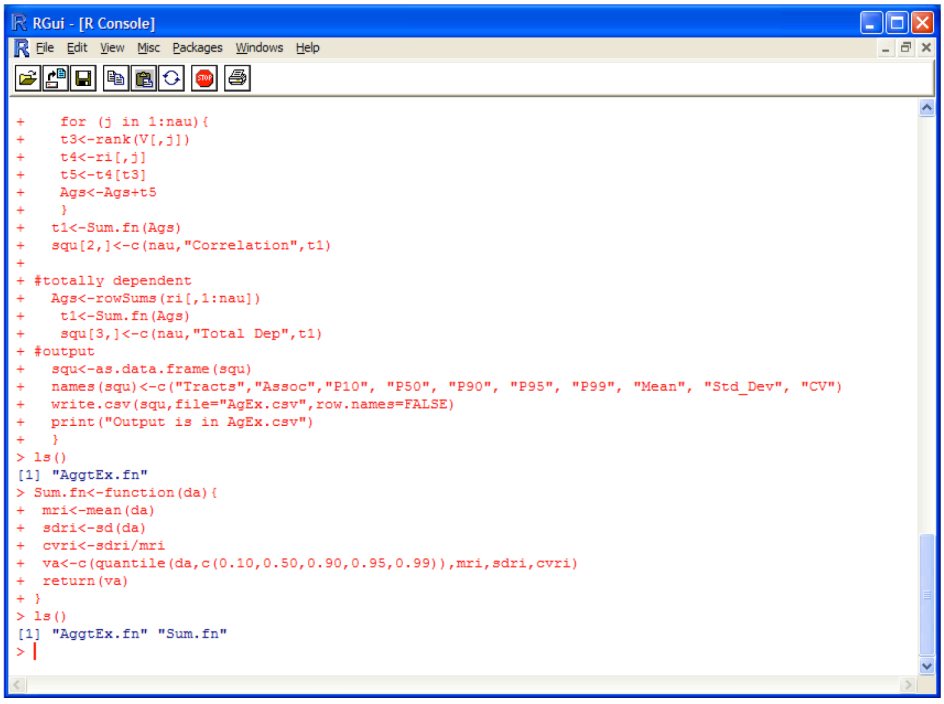
For purposes of illustration, let us call the R-folder ChuSarysu (the name is arbitrary). Let us name a folder as AggEx. This folder should contain the data files and ChuSarysu (the R-folder).

- Note that if you are opening R for the first time for this project, you will need to click on File, Change dir to change the directory to your training aggregation folder, which we have called AggtEx.
- Save workspace (File>SaveWorkspace). Call the workspace ChuSarysu in this example.

- Open your R project, copy and paste the functions in appendices 3 and 4.
- When in R, type ls() to see the functions that you have just pasted. They should be AggtEx.fn, and Sum.fn.

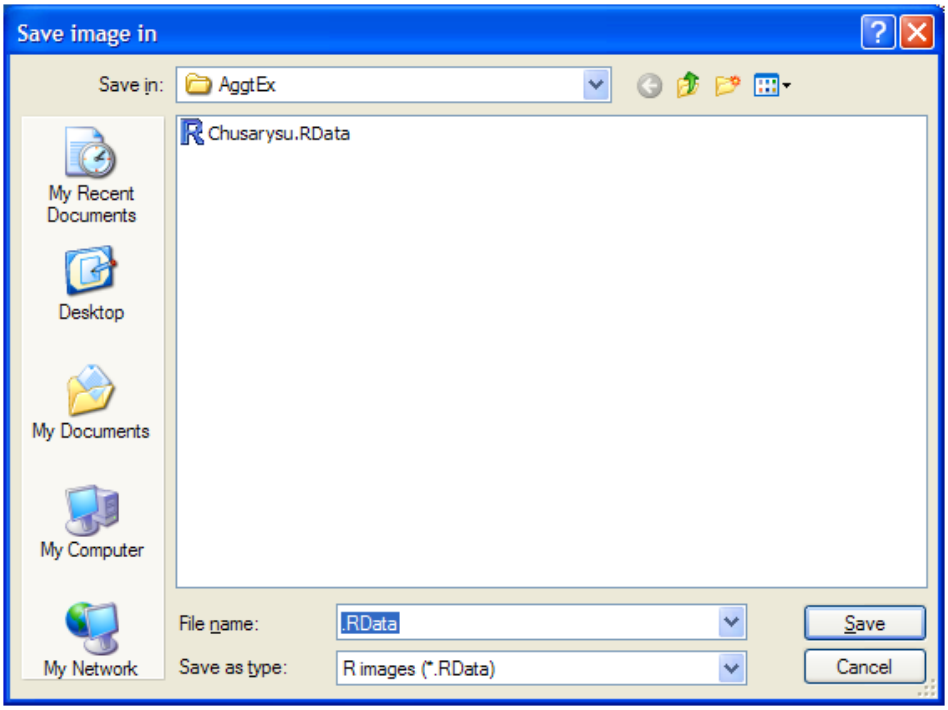


- Save Workspace (File>Save Workspace…). You should see a window called “Save image in”. At the bottom you will see “File name:” and “Save as type:” You should see an R folder (big blue R) in your folder unless this is the first time you have used and saved an R folder. You can either replace your existing R folder or provide a new file name and click on save.



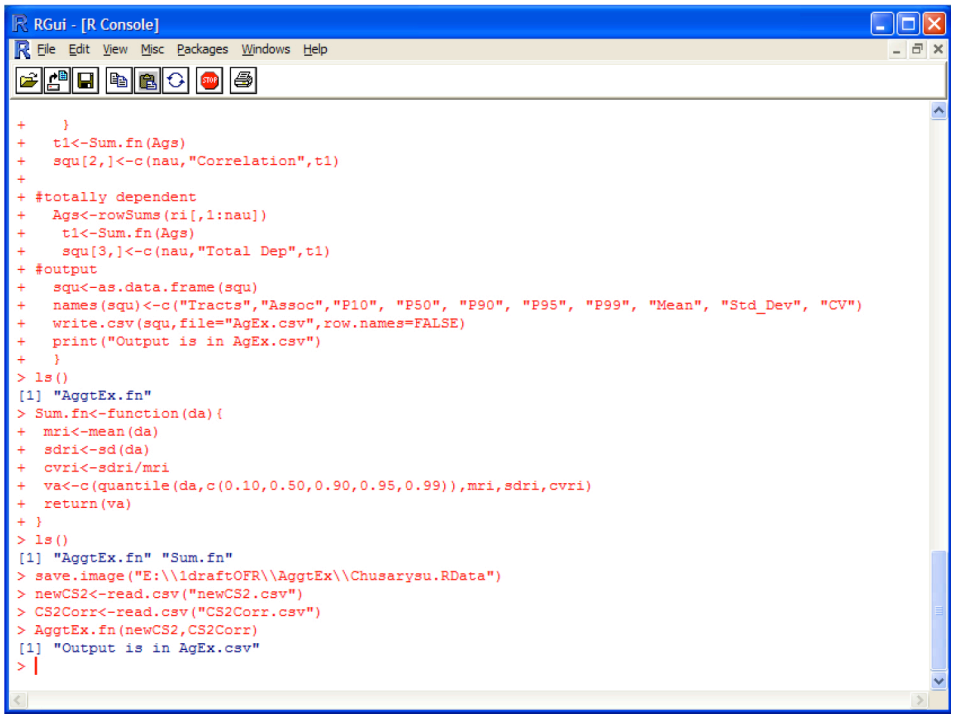
After preparing the input files, called newCS2.csv (appendix 1) and UsrCorr.csv (appendix 2) in this example, then do the following.

- When in your R folder, execute the following command to read in newCS2.csv:
- newCS2<-read.csv("newCS2.csv").
- Then execute the following command to read in UsrCorr.csv:
- UsrC<-read.csv("UsrCorr.csv",row.names=1).
- These two files are now in your R folder named newCS2 and UsrC respectively. Again, you can name them anything. Saving the worksheet will retain them in this folder.
- To generate output execute:
- AggtEx.fn(newCS2,UsrC).
- The output file AgEx.csv is written to your folder (say AggEx).



- At this point you should left click on File and Save Workspace.

Note that when AggEx.fn writes the function AgEx.csv to your user folder, it will replace a file by that name.



Produced in Menlo Park, California
Manuscript approved for publication April 27, 2011
Edited by Tracy Suzuki
Layout and design by Jeanne S. DiLeo