Kolmogorov Smirnov (Max-KS) in Banking Credit Risk Data Quality Control

Mark An Ph.D.

Credit Risk Analytics, Risk Management

CIBC

May 2010

Business Issue

Concern: Monitoring data quality during data transfer (download of external data or upload internal data to server)

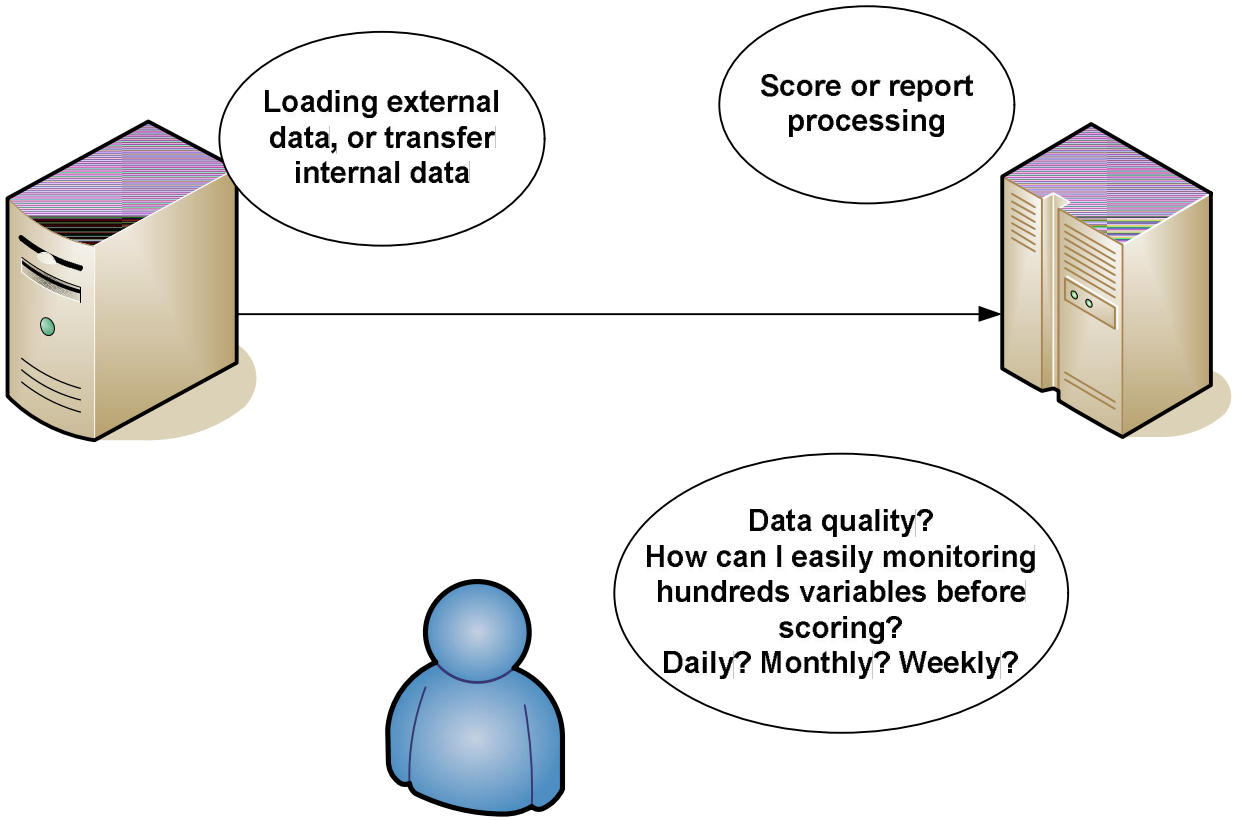
Solution:

- Automatically running SAS program in UNIX environment
- Email results to team

Benefits:

- Decreases labour cost
- Eliminates human error risk

Processing


Loading external data, or transfer internal data
Score or report processing
Data quality?
How can I easily monitoring hundreds variables before scoring?
Daily? Monthly? Weekly?

Case Study 1-(Daily)

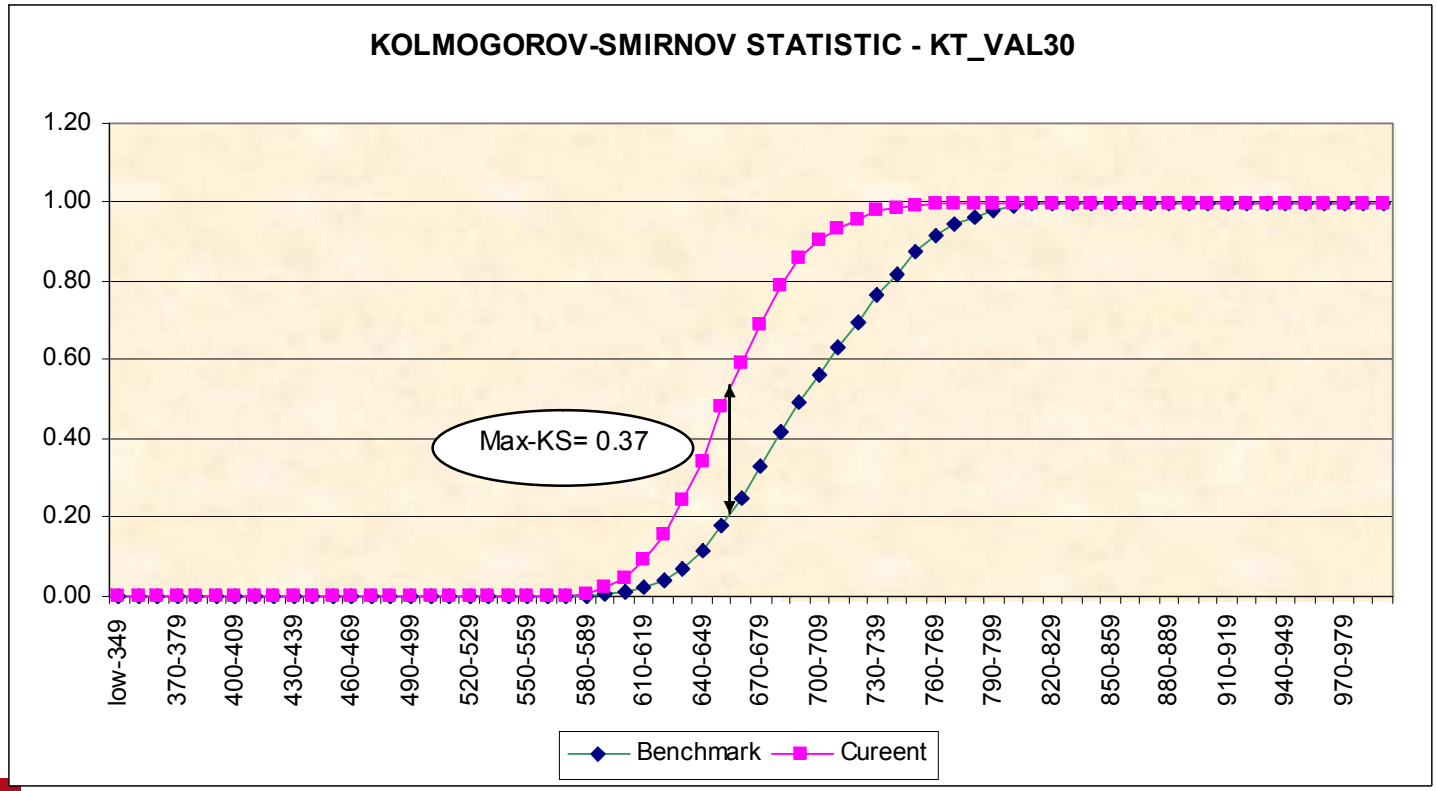
- A Bank can have a daily external data transfer process of over 10,000 records with 200 Bureau variables
- Monthly report shows scores shift
- After investigation, a SAS code error is found in column input step from Bureau's data to bank's datasets
- Result: Score calculations are wrong

Case study 2-(Monthly)

- **Monthly**, a bank's account management team need pulls data from a link file containing over 700 variables and millions of records, to create a Triad month dataset with 250 variables and millions of records
- Data quality control is required before monthly reporting

Introduction to Max-KS

Kolmogorov-Smirnov(KS) measures Maximum vertical separation (deviations) between two cumulative distributions (good and bad) in scorecard modeling



The Kolmogorov–Smirnov Test

Two-sample Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test can test whether two underlying one-dimensional probability distributions differ. The Kolmogorov–Smirnov statistic is

$$R_{n,n'} = \underset{x}{SUP}\left|F_n(X) - F_{n'}(X)\right|$$

H0: The data follow a specified distribution

Ha: The data do not follow the specified distribution Test

The two-sample test checks whether the two data samples come from the same distribution.

Limitations of KS Methodology

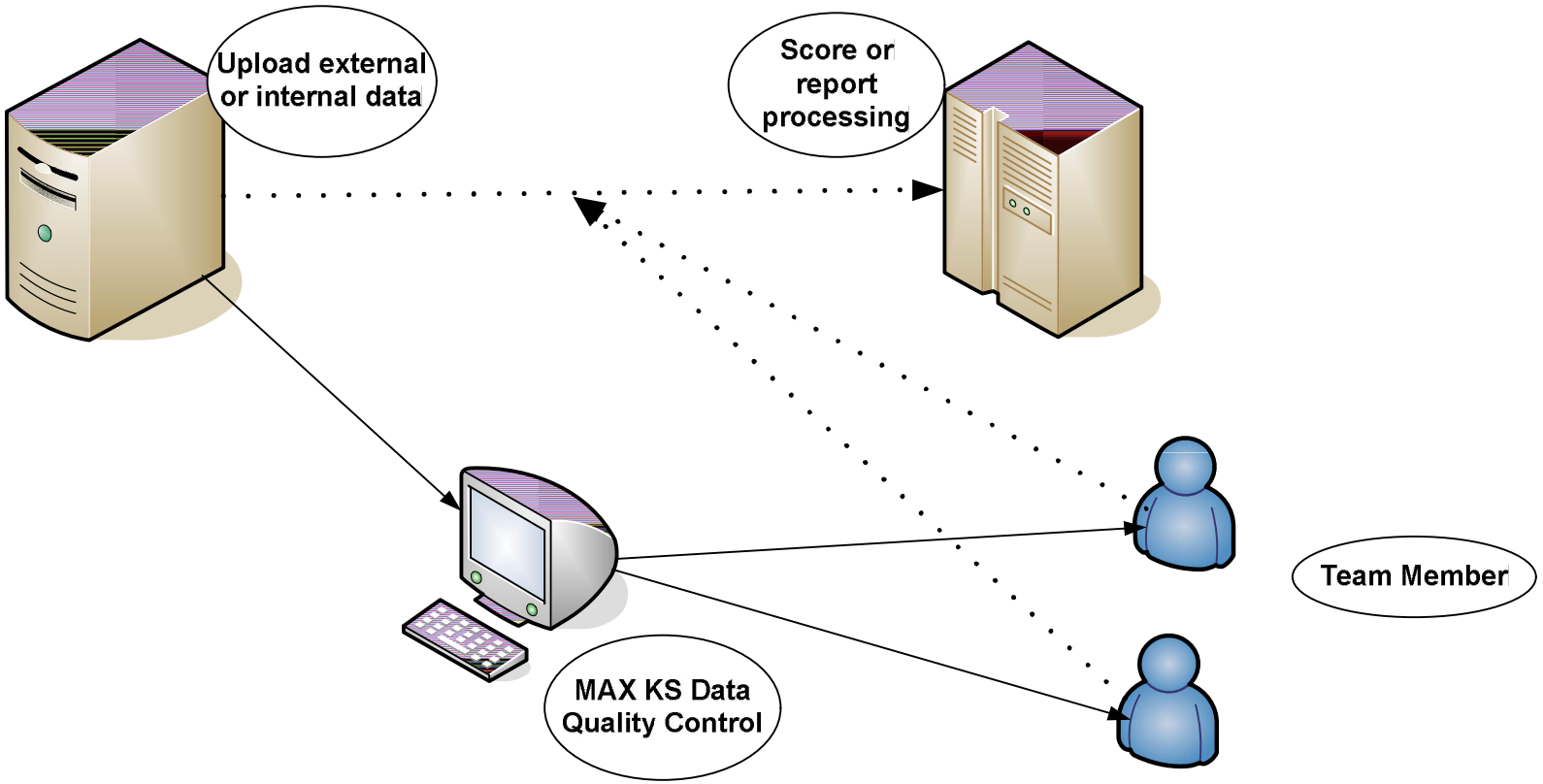
1. It is measured only at one point between two cumulated distributions, not along the entire range

2. Changes in the bin or bucket can change the Max KS Value

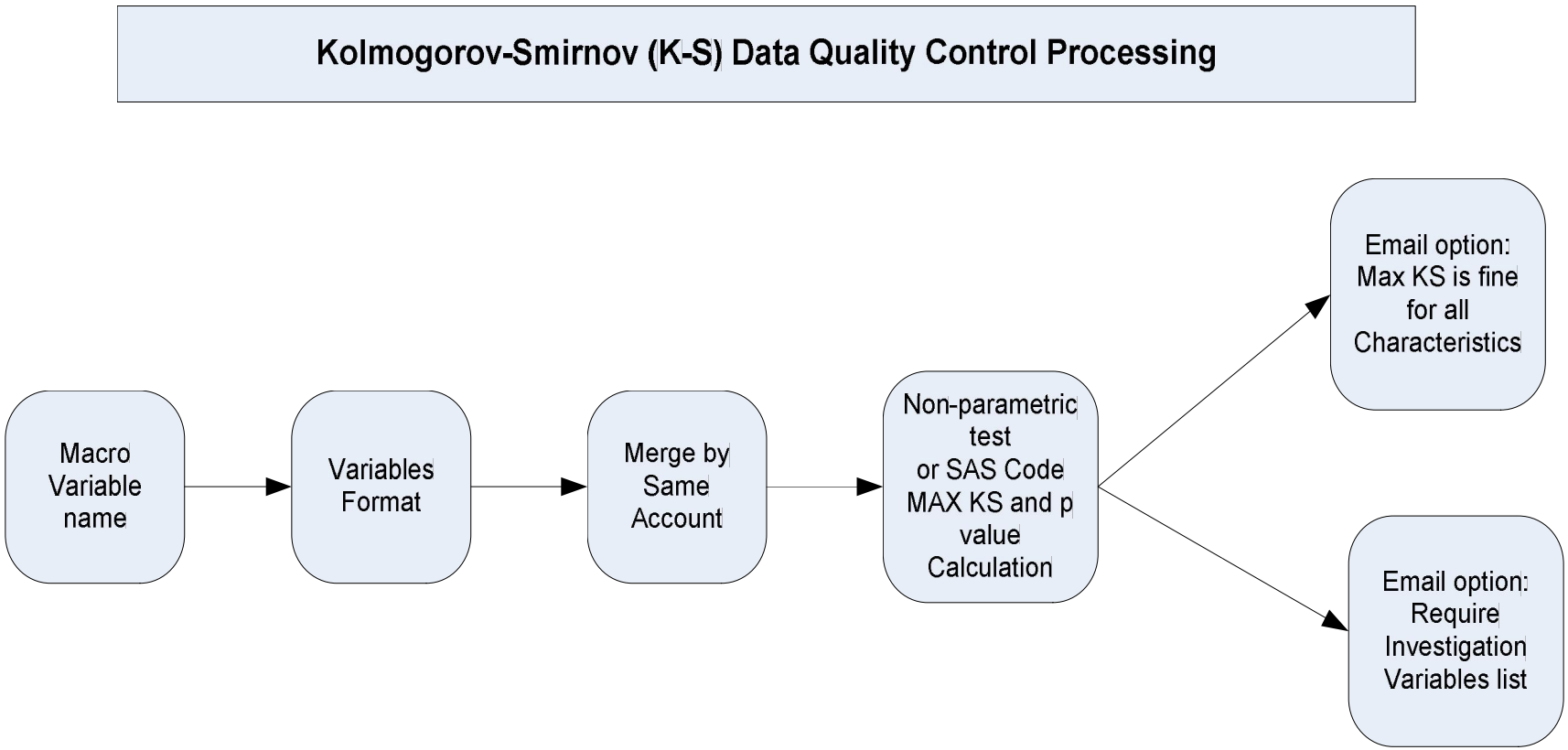
Details of KS Methodology

- Assumptions: benchmark data of variables values is correct
- Task: To evaluate current data values
- If p-value<0.05 or Max-KS is above the criterion, Max-KS test statistic rejects Null hypothesis
- Therefore, accept alternative hypothesis: current data values do not follow the distribution of benchmark data
- Conclusion: characteristic needs to be investigated (see example-Variable KT_VAL30).

Processing


Upload external or internal data
Score or report processing
Team Member
MAX KS Data Quality Control

Processing


Kolmogorov-Smirnov (K-S) Data Quality Control Processing
Macro Variable name
Variables Format
Merge by Same Account
Non-parametric test or SAS Code MAX KS and p value Calculation
Email option: Max KS is fine for all Characteristics
Email option: Require Investigation Variables list

SAS

Get variable name list

```
proc contents data=  out= name
(keep=name);
run;
```

Get macro variables:

```
proc sql noprint;
  select name into : name separated by ' '
from name;
quit;
```

Format all variables

(One variable for numeric, one for character)

```
proc format;
 value AT
 .     ='Missing'
 Low- -1 =' -1'
 0 - 5 = '0-5’
 …….
 116-120 = '116-120'
 121-125 = '121-125'
 126- High = '126+';

 value $GO
 'Y'='Yes'
 ' '=  'No'
 other='Check';

 run;
```

SAS

```
proc npar1way  edf data=a noprint  ;
     class Treatment;
     var Response;
     freq Freq;
     output out=KolSmir2Stats(keep=_D_ P_KSA);
  run;
```

Note:

_D_: Two-sample Kolmogorov-Smirnov Statistic

P_KSA: P-value, Two-sample Kolmogorov-Smirnov

SAS

```
proc tabulate data=final  missing out=a1(keep= &name._BM N);
        format &name._BM &name2..;
        class &name._BM / PRELOADFMT;
        title "Benchmark";
        table &name._BM ALL, N PCTN /printmiss misstext='0';
   run;
proc tabulate data= final  missing out=a2(keep= &name._CU N);
        format &name._CU &name2..;
         class &name._CU / PRELOADFMT;
         title "Current”;
          table &name._CU ALL,  N PCTN /printmiss misstext='0';
   run;
```

SAS

```
data MEG;
                set KolSmir2Stats;
                KS=round(_D_,.0001);
                p=round(P_KSA,.0001);
                *if p <=0.05
               *if ks>=0.10;
              / *If KT_val18 ks>=0.15*/
run;
```

Using UNIX, program at a specific date with specific conditions (IF find 20090228 data exist) running

```
if [ -f 20090228data ]
 then
  echo "run aa.sas"
  aa.sas
fi
```

SAS

```
%macro roll();
 %let file_flag = 0;

data _null_;
            set final;
                 if OBS =0 then call symput('file_flag','1');
                                  if OBS >0 then call symput('file_flag','2'); run;

  %if &file_flag = 1  %then %do;
   FILENAME mailx EMAIL "Mark.An@cibc.com" emailsys=SMTP
        SUBJECT="..... data &DTT. Max KS is fine for all Characteristics"
                                  CC=("A.A@cibc.com"  "B.B@cibc.com"  "C.C@cibc.com") ;
    %end;

   %if &file_flag = 2 %then %do;
            FILENAME mailx EMAIL "Mark.An@cibc.com" emailsys=SMTP
                 SUBJECT="..... for &DTT. Require Investigation"
                                  ATTACH="/...../Data Monitoring/Variable Distribution/u_&dtt..pdf"
                                                CC=("A.A@cibc.com"  "B.B@cibc.com"  "C.C@cibc.com") ;
  DATA email;  set final;
                 FILE mailx;   put ;
                                  put " name= " name "Description=" Description  "Max KS = " KS;  Run;
   %end;
%mend;
```

Example

Value	# of Error data	% of error data	Cumulative% of Error data	# of Benchmark	% of Benchmark	Cumulative% of Benchmark	MAX-KS
Missing	0	0	0	0	0	0	0
< -1	0	0	0	0	0	0	0
0-5	0	0	0	17053	26.4%	26%	26.4%
6-10	0	0	0	15699	24.3%	51%	50.7%
11-15	0	0	0	12174	18.9%	70%	69.6%
16-20	0	0	0	8272	12.8%	82%	82.4%
21-25	0	0	0	5258	8.1%	91%	90.5%
26-30	0	0	0	2912	4.5%	95%	95.0%
31-35	0	0	0	1573	2.4%	98%	97.5%
36-40	0	0	0	828	1.3%	99%	98.8%
41-45	0	0	0	409	0.6%	99%	99.4%
46-50	0	0	0	202	0.3%	100%	99.7%
51-55	0	0	0	97	0.2%	100%	99.8%
56-60	0	0	0	44	0.1%	100%	99.9%
61-65	0	0	0	23	0.0%	100%	100.0%
66-70	0	0	0	12	0.0%	100%	100.0%
71-75	0	0	0	11	0.0%	100%	100.0%
76-80	0	0	0	4	0.0%	100%	100.0%
81-85	0	0	0	2	0.0%	100%	100.0%
86-90	0	0	0	1	0.0%	100%	100.0%
91-95	0	0	0	1	0.0%	100%	100.0%
96-100	456	0.7%	0.7%	0	0.0%	100%	99.3%
101-105	589	0.9%	1.6%	0	0.0%	100%	98.4%
106-110	245	0.4%	2.0%	0	0.0%	100%	98.0%
111-115	123	0.2%	2.2%	0	0.0%	100%	97.8%
116-120	5653	8.8%	10.9%	0	0.0%	100%	89.1%
121-125	8356	12.9%	23.9%	0	0.0%	100%	76.1%
126+	49,153	76.1%	100.0%	0	0	100%	0.0%
Max KS	64,575			64,575			1

Conclusion

• The K-S statistic is very useful during data quality control. The assumption must be made that the benchmark data is correct.

• If the current data variable distribution follows the benchmark distribution, we can safely say, that the current data is correct.

• If Max-KS is above the criterion, which is set based on the objective and data, the program will automatically list need investigate variables, send an email to team and decide whether or not to stop score processing.

Questions?

References

- *Edward M. Lewis An Introduction to Credit Scoring*
- *Christopher M. Bishop Pattern recognition and Machine learning*
- *An Overview of Non-parametric Tests in SAS:When, Why, How*
- *Derek Montrichard, Reject Inference Methodologies in Credit Risk Modeling*
- *Erich, Joseph P. Romano Testing statistical hypotheses*
- Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*
- Jerome Friedman Trevor Hastie Robert Tibshirani The Elements of Statistical Learning
- Charles T. Clark & Lawrence L. Schkade Statistical analysis for Administrative Decision