

# Global Credit Data

*by banks for banks*

## Proving Representativeness in your GCD Samples



Ben Galow and Nina Brumma, GCD

# AGENDA

- ❑ Motivation & Overview
- ❑ What do other banks do? Methcom Survey results
- ❑ Examples

**Anti Trust Warning:** Participants are warned not to provide sensitive information about their bank or to engage in discussions which might encourage or lead to collusive behaviour. If in doubt then please seek guidance from your own bank's policies or legal counsel.

**Disclaimer:** Any views expressed in this presentation are those of the presenter and do not necessarily represent the views of Global Credit Data or its members.



## Your Input is required

Discussion elements and questions for today

# Motivation

- When it comes to using GCD (or any pooled) data two items are asked immediately

## What about Data Quality?

→Data Quality Framework & Dashboard



Share your experience

## How can we prove representativeness?

“A general concern is the representativeness and how we can prove that (e.g. new DoD). We are considering using data for CRE, but the representativeness was raised as a concern by the Modelling team.”

“Even for benchmarking, the regulatory requirement on representativeness of the data has to be applied. Sure, we can slice and dice the data. However, the underlying process on the data collection from each of the member banks is unclear if not unified”

(citations from 2022 GCD data usage survey)

# What is representativeness?

## Regulatory purpose

eg. SR 11/7 or EBA GL  
„institutions should have [...] methods for assessing the representativeness of data used for the purpose of estimation of risk parameters“

## Analytical purpose

A representative sample is a subset of a population that seeks to accurately reflect the characteristics of application portfolio

## Application purpose

The type of application defines what representativeness means and necessitates in terms of analyses

# Regulatory Overview – EU EBA GL 2017/16

- ❑ Guidelines include 5 **dimension** of assessment of representativeness
- ❑ Explicit **qualitative** and **quantitative** elements
- ❑ Use of external and pooled data explicitly mentioned:  
*“same standards and methods for representativeness of data stemming from different sources (incl. external/ **pooled** data)”*



Dimension	Model Development	Model Calibration
Scope of application ( § 29)	X	X
Definition of default ( § 30)	X	X
Distribution of the relevant risk characteristics ( § 31)	X	X
Lending standards and recovery policies ( § 33)	X	X
Current and foreseeable economic or market conditions ( § 32)	-	X

# What is needed and how GCD can assist



## Framework / „cooking recipe“

- ❑ A document describing the steps a member bank should do when assessing representativeness
- ❑ The following questions shall be answered
  - What does representativeness mean?
  - What are the regulatory requirements? (Focus on EU and US)
  - Which elements must be included?
  - Which quantitative and qualitative analysis must be included?
  - How can you come to a final answer if data is useful for your purpose?
  - How can you overcome issues (MoC)?



## Examples for different facility asset classes

- ❑ Proof-of-Concept document
- ❑ Definition/ derivation of suitable RDS
- ❑ Qualitative and quantitative analyses structured along regulatory dimensions
  - the scope of application
  - the definition of default
  - the distribution of the relevant risk characteristics
  - lending standards and recovery policies
  - the current and foreseeable economic or market conditions



## Roles with GCD and members

- ❑ GCD:
  - Providing / maintaining framework
  - Guidance to members
  - Running a pilot with selected members
  - Conducting analyses on pool level
  - ...
- ❑ Members:
  - Applying framework
  - Conducting analyses on application portfolio
  - Providing feedback to GCD
  - ...

# What do other banks do?

## Methcom Survey Results

### 1. Scope of representativeness framework

#### a. What dimensions are included in your representativeness framework?

i. Scope of application		5
ii. Definition of default		4
iii. Distribution of the relevant risk characteristics		6
iv. Current and foreseeable economic or market conditions		3
v. Lending standards and recovery policies		3

#### b. Do you distinguish between elements for representativeness analysis for model RDS and calibration RDS?

yes		3
no		2



Share your experience

# What do other banks do?

## Methcom Survey Results

### 2. Quantitative Methods

#### a. Which method does your bank use for performing distributional comparison of risk drivers?

i. Hellinger distance score		1
ii. Population Stability Index Test		3
iii. Kolmogorow-Smirnow Test		1
iv. Graphical visualization of distribution of risk drivers over time		3
v. Other. Which one?		2

#### Other:

- Hellinger distance we use for categorical risk drivers, Wasserstein-distance we use for metric risk drivers. Each with a threshold of 25%.
- Value Range Test and Binomial Test & Chi-Squared

#### b. Which are key risk drivers that you use in your quantitative representativeness analyses for LGD modelling?

i. Time to resolution		2
ii. Time to peak recovery		1
iii. Seniority code		3
iv. Collateral indicator		4
v. Guarantee indicator		4
vi. Cure		1
vii. Other. Which?		4

#### Other:

Geography, Industry, Financial Metrics, Type of Collateral, EAD, Segments

# What do other banks do?

## Methcom Survey Results

### 3. Internal processes

a. At which stage do you use representativeness analyses?

i. Model Development		1
ii. Calibration		0
iii. Both		4

b. At which frequency do you perform representativeness analyses?

i. Within validation cycle		3
ii. Other. Which?		2

Other:

Annual model performance/validation review

During model redevelopment which is around every 5-7 years and during quarterly monitoring..

c. Do you use the same framework/methods for internal and external data?

i. Yes		5
ii. No		0

d. How do you deal with changes in processes, e.g., definition of default?

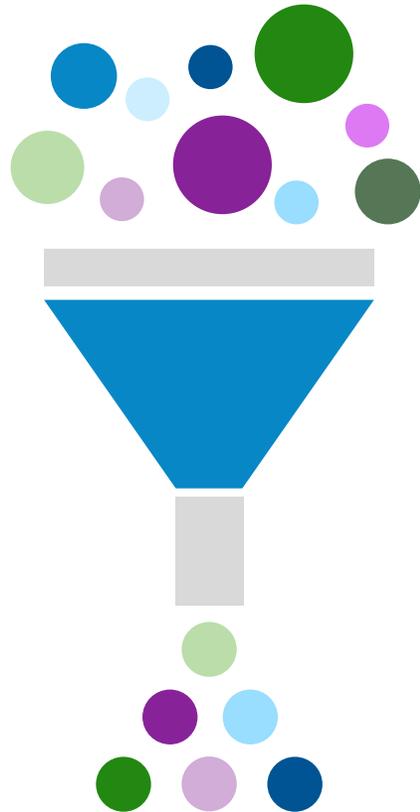
i. Roll-out current definition to historical cases?		3
ii. Introduce Adjustment/MoC		4
iii. Other. What?		1

Other:

Nothing. We have not considered this in our usage of GCD data.

# From raw to usable data: Start with Reference Data Set

## GCD's Standard RDS



- Reference Data set (RDS) refers to the data set after application of filters which is used for analysis.
- The RDS is the raw data qualification and is the first step in the representativeness analysis.

Filter	Stage	Unresolved	Year of Default Small	Default	Amount Incomplete Portfolio	Validation Rules	Nr of loans	Total
Raw Data Set w/o Filters	initial	3,359	2,443	7,003	1,868	2,066	6,912	37,838
Keep resolved defaults	before	3,359	2,443	7,003	1,868	2,066	6,912	37,838
	after	0	1,774	6,138	1,755	1,579	6,108	34,479
Keep years of default 2000-2018	before	0	1,774	6,138	1,755	1,579	6,108	34,479
	after	0	0	5,920	1,755	1,462	6,037	32,705
Exclude small default amounts	before	0	0	5,920	1,755	1,462	6,037	32,705
	after	0	0	0	793	663	3,958	26,785
Exclude incomplete portfolio data	before	0	0	0	793	663	3,958	26,785
	after	0	0	0	0	525	3,614	25,992
Exclude older data not in line with latest data VRs	before	0	0	0	0	525	3,614	25,992
	after	0	0	0	0	0	3,491	25,467
Minimize facility weighting effects	before	0	0	0	0	0	3,491	25,467
	after	0	0	0	0	0	0	0
Reference Data Set	final	0	0	0	0	0	0	<b>21976</b>

# (new) Definition of Default

2016

EBA publishes the finalized GL on the application of the definition of default past due (EBA/GL/2016/07)

2018

Application for banks under the SSM (two-step approach)

2020

Until end of 2020 new DoD should be operationalized

2021

New DoD basis for reporting; Models need to be updated according to new DoD under IRB repair



- What did you do? What have you done w.r.t. historical data remediation?
- Did it affect GCD submission?
- What do you need from GCD to prove DoD is ok?

# Distribution of relevant risk characteristics

## Identify key risk drivers for quantitative distributional comparison:

### Security information:

- Collateral indicator: Indicates loan has underlying protection in the form of collateral or security.
- Guarantor indicator: Indicates loan has underlying protection in the form of a guarantee, a CDS or some support from a Key party

### LGD Scenario information:

- Cure indicator: Indicator to show if the loan is cured or not

### Recovery duration information:

- Time to resolution: Number of Days between default and resolution date of the loan
- Time to recovery: Average number of days for recovery transactions to come in after default

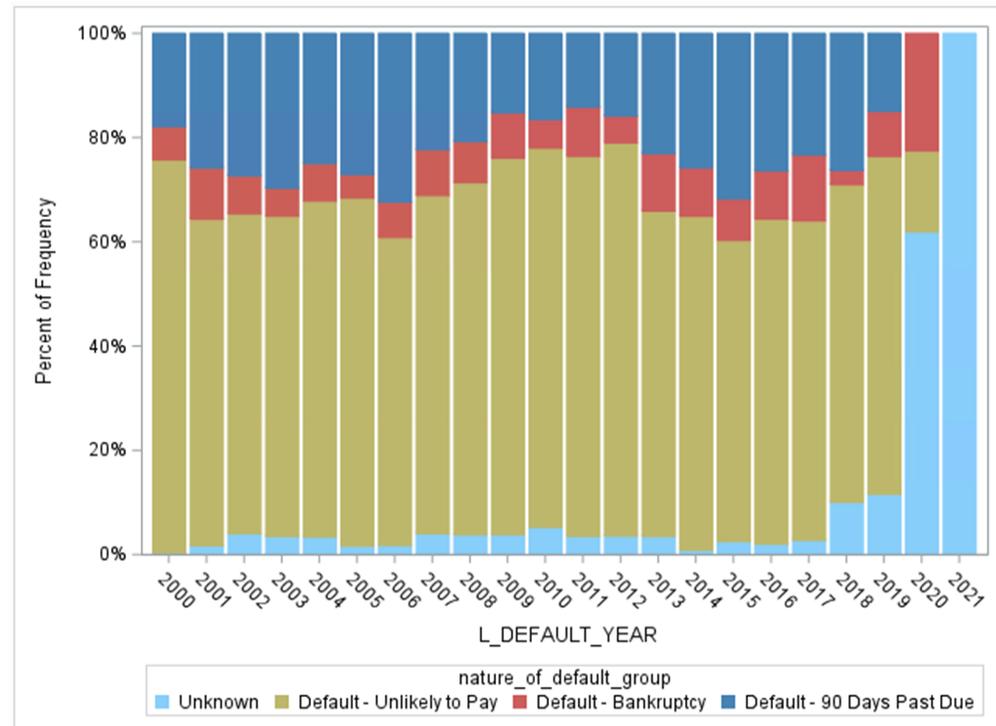
### Contractual details:

- Seniority Code

# Example – Changes over time analysis

## Qualitative analysis for Definition of Default:

Banks are required to show that the definition of default is consistent over time and different jurisdictions. GCD uses the Basel definition of default.



- Distribution over time looks relatively stable
- There is a visible trend starting around the financial crisis years in 2008: increase of Unlikely to pay and decrease of DPD90-events.
- In most recent years there is a change in distribution which can be attributed to a low number of total data points and an overrepresentation of quickly resolving cases (resolution bias).

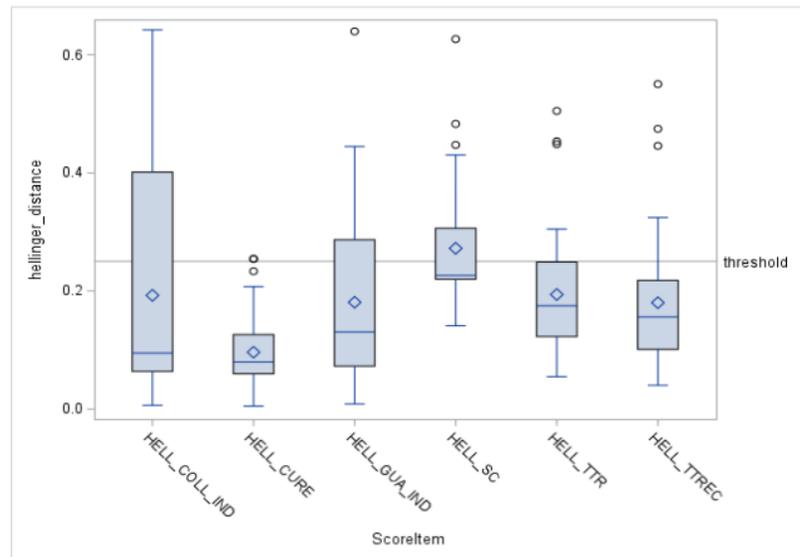


Share your experience

# Example – Distribution check with Hellinger Distance

## Quantitative assessment of representativeness:

- ❑ Leverage the already approved distributional comparison test based on Hellinger distance measure from the DQ Dashboard (**pass threshold: Hellinger distance < 0.25**)
- ❑ Compare the similarity of each individual bank to the overall data pool using the Hellinger distance measure w.r.t. to the key risk drivers performed on the afore defined samples.



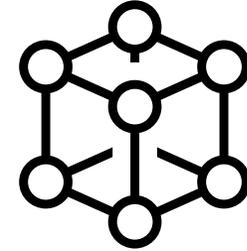
Which metrics do you use?

For most of the member banks the **risk driver** profiles show **similar pattern on the afore defined RDS**.

# What does GCD plan to do?



Want to be involved? Reach out to  
[Nina.brumma@globalcreditdata.org](mailto:Nina.brumma@globalcreditdata.org)



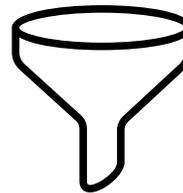
Include results from  
representativeness  
analyses to refine RDS

**FURTHER ANALYSES**

**RDS REFINEMENT**

**FRAMEWORK**

Analyse further  
dimensions like scope  
of application,  
definition of default...



Develop and provide  
standard framework  
to member banks

# Appendix: Hellinger distance

## Hellinger distance measure

- ❑ Suppose you have two discrete probability distributions  $P=(p_1,p_2,\dots,p_k)$  and  $Q=(q_1,q_2,\dots,q_k)$  with relative frequencies  $p_i,q_i$  for the possible realizations  $i=1,2,\dots,k$ .
- ❑ Test if the probability distribution of the dataset used for model calibration (pool),  $P$ , is representative for the probability distribution  $Q$  of the application portfolio (lender portfolio)

- ❑ The Hellinger Distance is defined as 
$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{p_i} - \sqrt{q_i})^2}$$

- ❑ It gives values between 0 and 1, with  $H=0$  meaning that both distributions are identical and  $H = 1$  meaning that they are singular.
- ❑ To demonstrate representativeness values close to 0 are desirable, with values above 0.25 considered critical.