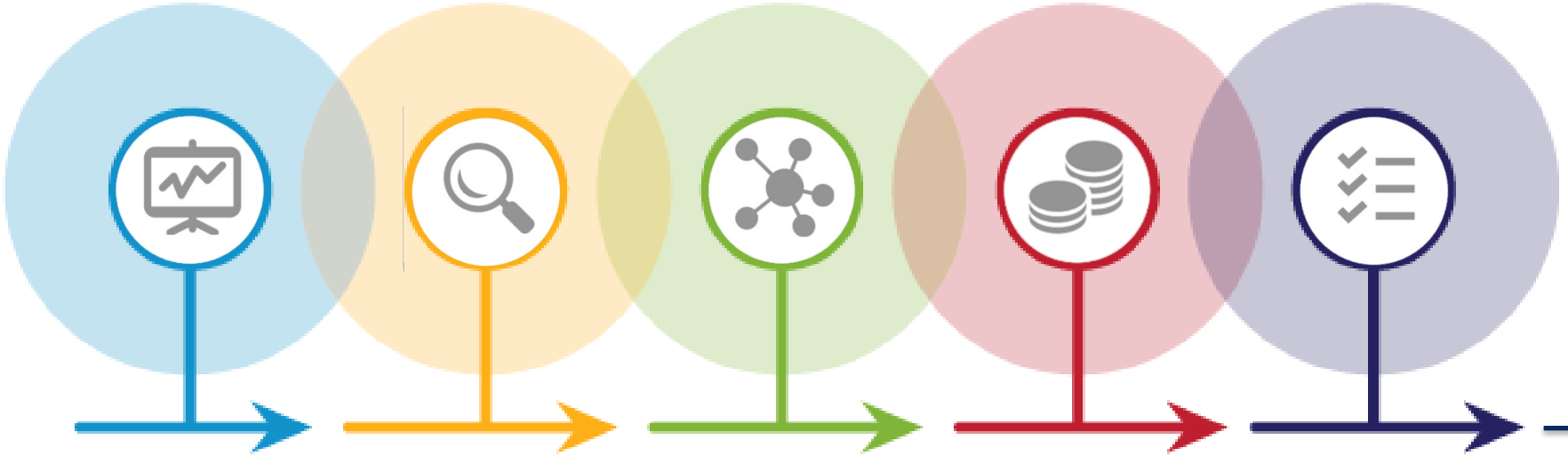
DAGLI STUDI DI SETTORE ALL' INDICE SINTETICO DI AFFIDABILITÀ

EVOLUZIONE E INNOVAZIONE DELLA METODOLOGIA

ROMA, 4 MARZO

RELATORI: ARIANNA CAMPAGNA E GIANCARLO FERRARA

Il Processo metodologico



1. ANALISI DELLA QUALITÀ DEI DATI

2. INDIVIDUAZIONE DEI FATTORI CARATTERISTICI

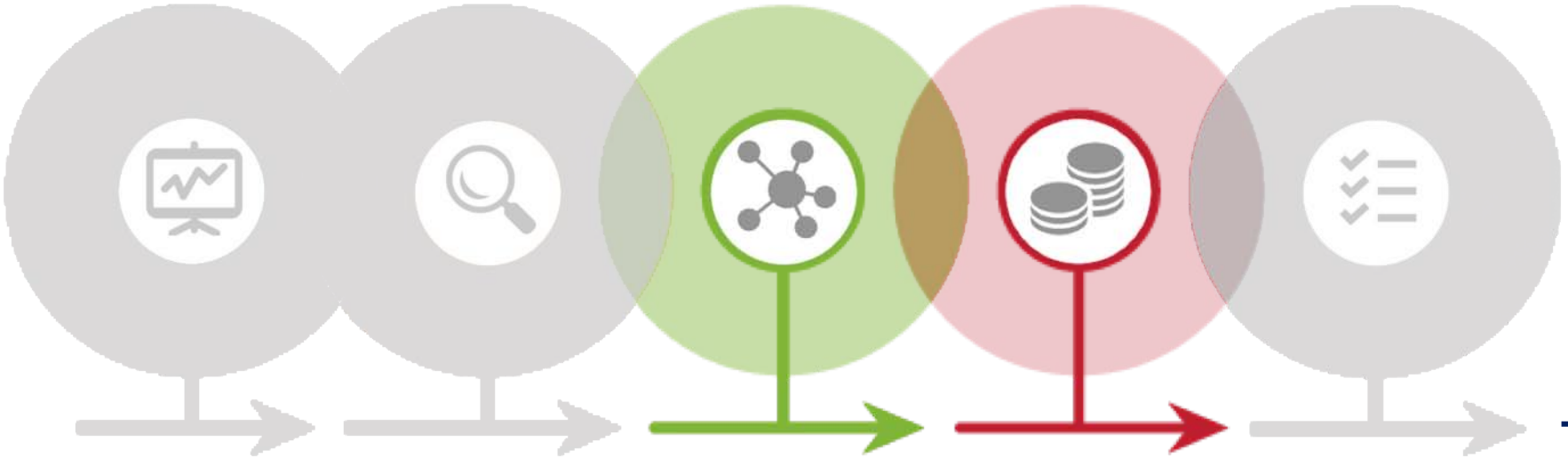
3. INDIVIDUAZIONE DEI GRUPPI OMOGENEI (ANALISI FATTORIALE E CLUSTER ANALYSIS)

4. STIMA DEI RICAVI /COMPENSI, DEL VALORE AGGIUNTO E DEL REDDITO

5. NUOVO SISTEMA DI INDICATORI ELEMENTARI
   - AFFIDABILITÀ
   - ANOMALIE

6. INDICE SINTETICO DI AFFIDABILITÀ

Focus



3. INDIVIDUAZIONE DEI GRUPPI OMOGENEI (ANALISI FATTORIALE E CLUSTER ANALYSIS)

4. STIMA DEI RICAVI /COMPENSI, DEL VALORE AGGIUNTO E DEL REDDITO

6. INDICE SINTETICO DI AFFIDABILITÀ

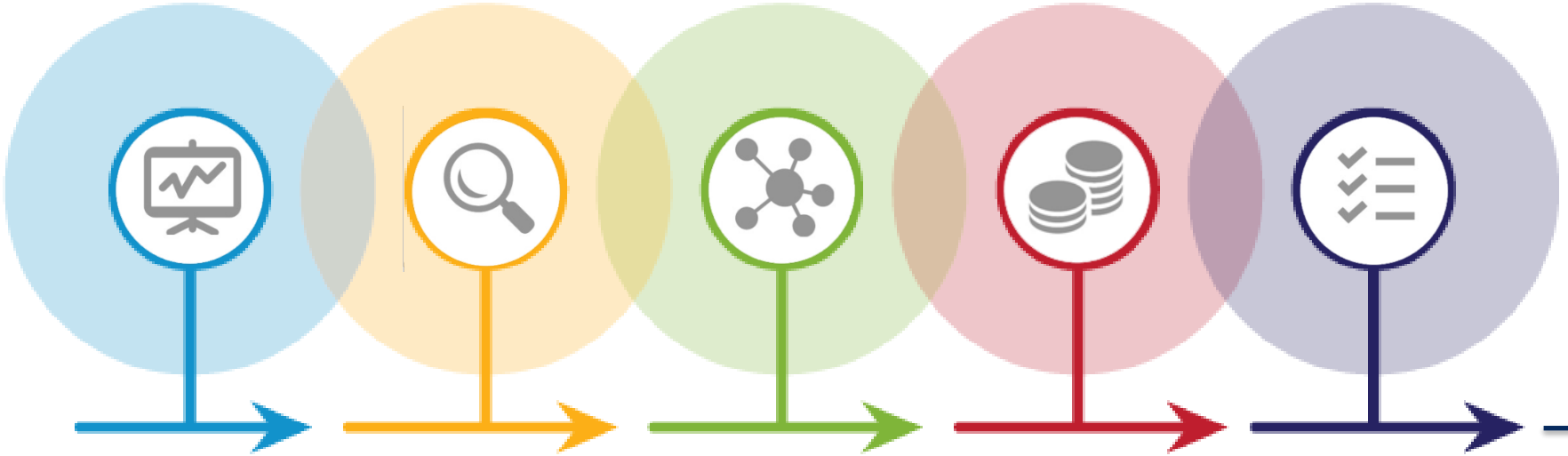
Obiettivo

- **Misurare l'affidabilità del contribuente**
- **Definizione di un valore di riferimento «normale» delle basi imponibili**
- **Stima delle basi imponibili**

Metodologie: aree di intervento

- ✓ Approccio economico alla base della **analisi dei gruppi omogenei**
- ✓ Approccio statistico nella stima dei gruppi
- ✓ Approccio econometrico nella definizione delle **funzioni di stima**

Il Processo metodologico



1. ANALISI DELLA QUALITÀ DEI DATI

2. INDIVIDUAZIONE DEI FATTORI CARATTERISTICI

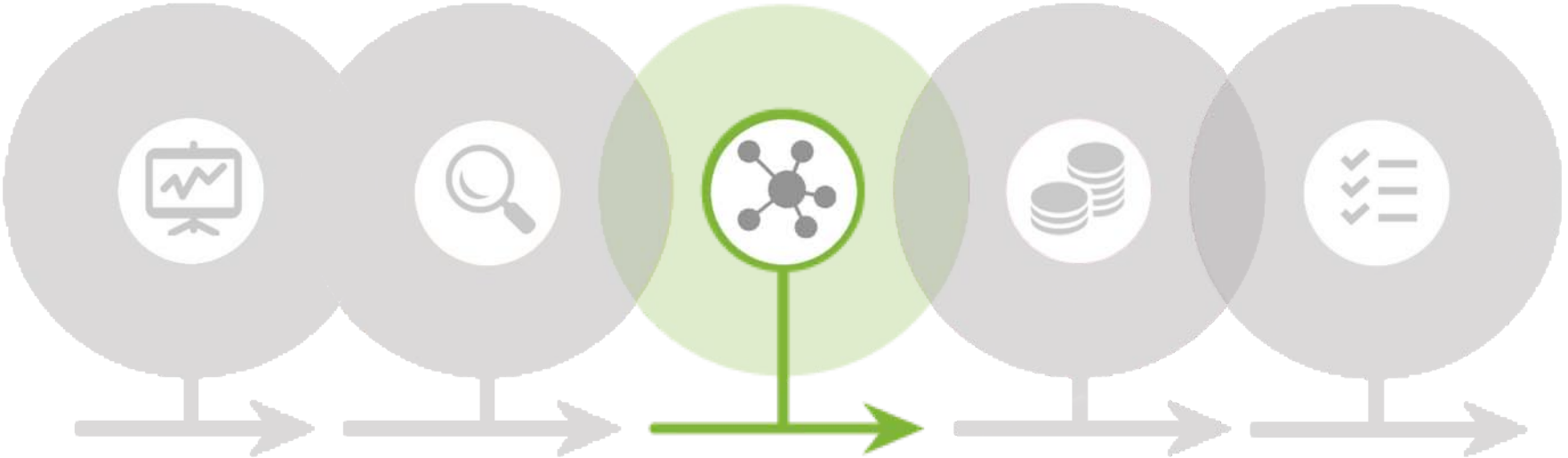
3. INDIVIDUAZIONE DEI GRUPPI OMOGENEI (ANALISI FATTORIALE E CLUSTER ANALYSIS)

4. STIMA DEI RICAVI /COMPENSI, DEL VALORE AGGIUNTO E DEL REDDITO

5. NUOVO SISTEMA DI INDICATORI ELEMENTARI
   - AFFIDABILITÀ
   - ANOMALIE

6. INDICE SINTETICO DI AFFIDABILITÀ

Cluster analysis



3. INDIVIDUAZIONE DEI GRUPPI OMOGENEI (ANALISI FATTORIALE E CLUSTER ANALYSIS)

6. INDICE SINTETICO DI AFFIDABILITÀ

Analisi dei gruppi: perché?

Eterogeneità indotta dal contesto

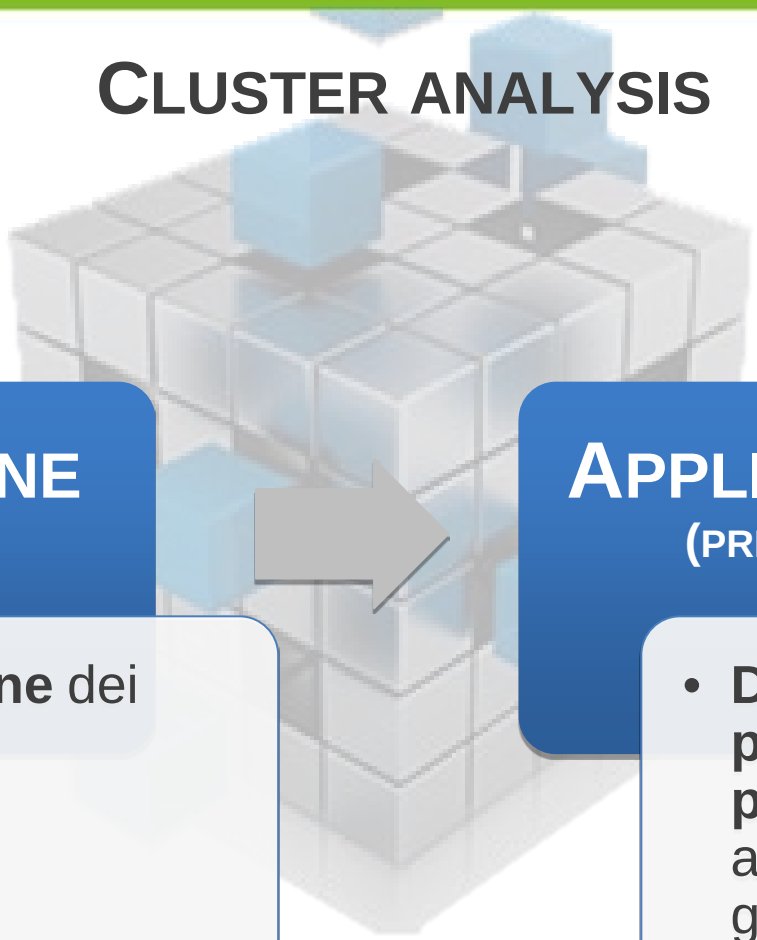
Contesto misurato attraverso lo studio delle modalità organizzative

Gruppi omogenei (cluster)

Variabilità non spiegata sintetizzata dalla variabile cluster

NECESSITÀ

CLUSTER ANALYSIS



COSTRUZIONE (STIMA)

- **Definizione** dei **gruppi**

APPLICAZIONE (PREDIZIONE)

- **Definizione delle probabilità a posteriori** di appartenenza ai gruppi

IDENTIFICAZIONE

ANALISI FATTORIALE

Con riferimento agli aspetti statistici associati alla teoria economica sottostante, i diversi blocchi logici alla base del metamodello vengono analizzati attraverso

singole Analisi fattoriali,

in modo da individuare coerentemente le diverse dimensioni di analisi utilizzate per la definizione dei gruppi.

‘‘Blocchi costitutivi’’

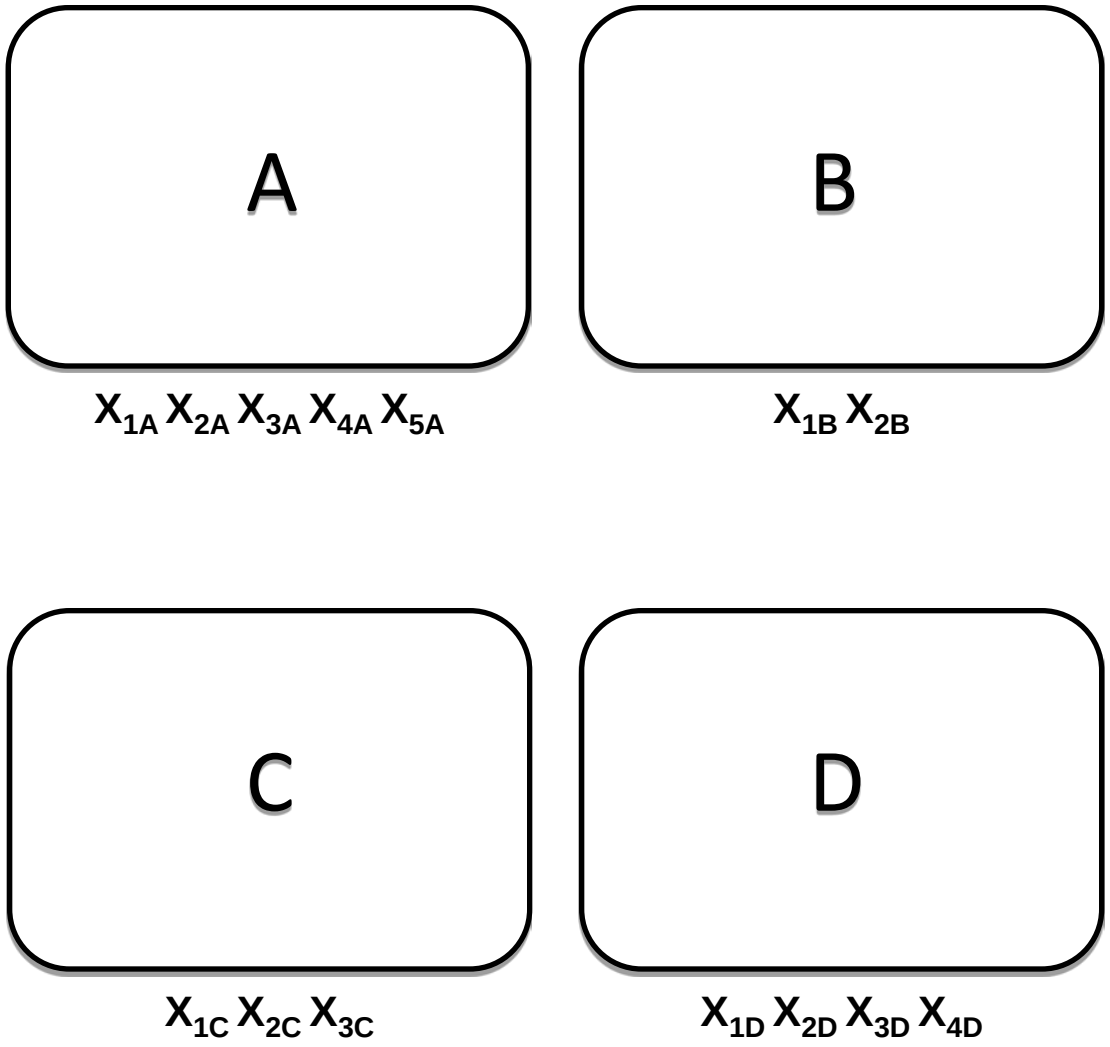
A

B

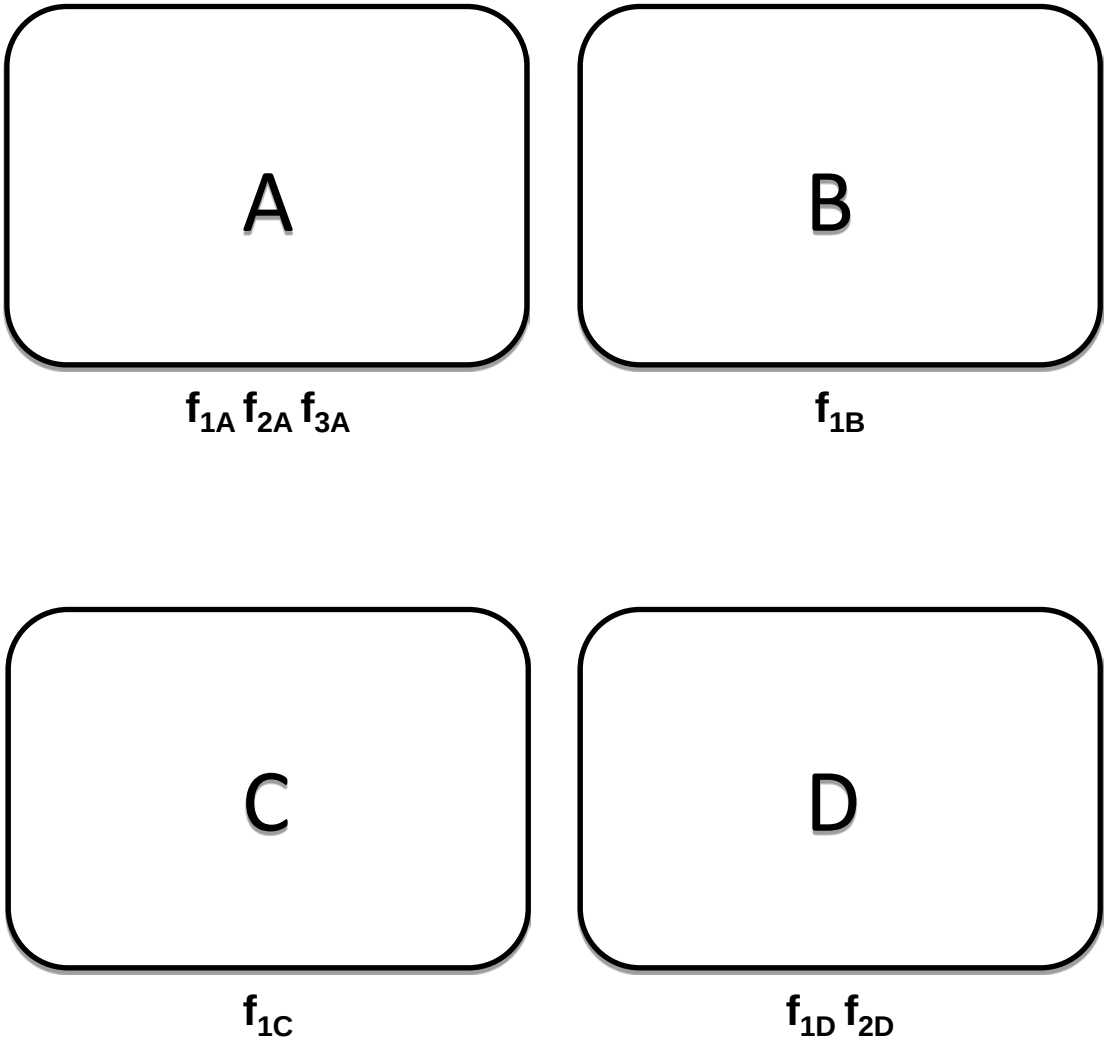
C

D

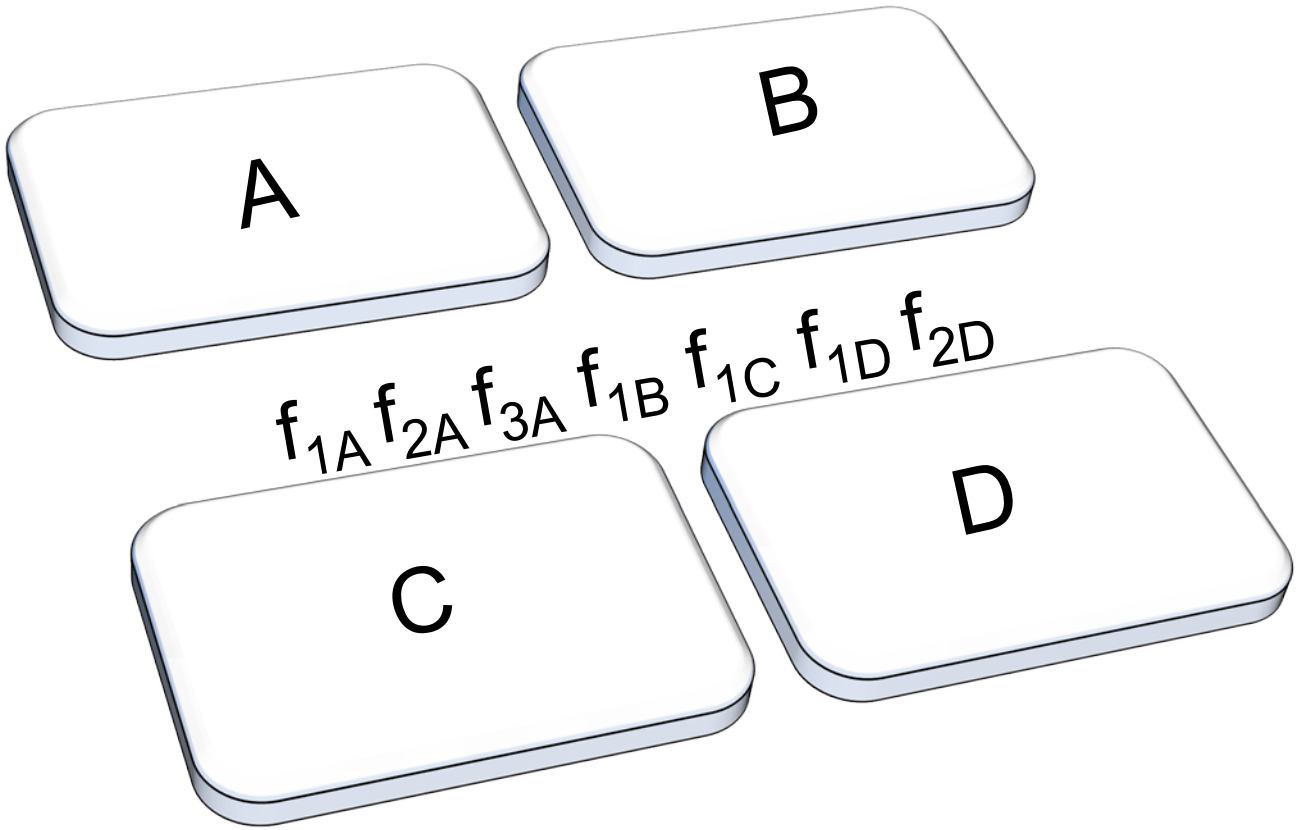
ANALISI FATTORIALE



ANALISI FATTORIALE



ANALISI FATTORIALE



FINITE MIXTURE

$$f(\boldsymbol{x}) = \sum_{g=1}^{G} p_g f_g(\boldsymbol{x})$$

dove:

- $f_g(\boldsymbol{x})$ sono funzioni (di densità) di probabilità (componenti ≡ sottopopolazioni) che possono assumere differenti forme analitiche (normale, Poisson, esponenziale, etc.);
- p_g (pesi o probabilità *a priori*).

$f(\boldsymbol{x})$ è una funzione di densità, ovvero non negativa e con integrale pari a 1 (Everitt, 1981).

Una osservazione $\boldsymbol{x}$ viene classificata nelle G sottopopolazioni attraverso il calcolo delle probabilità a posteriori (McLachlan and Peel, 2000):

$$p(g|\boldsymbol{x}) = \frac{p_g f_g(\boldsymbol{x})}{\sum_{h=1}^{G} p_h f_h(\boldsymbol{x})}$$

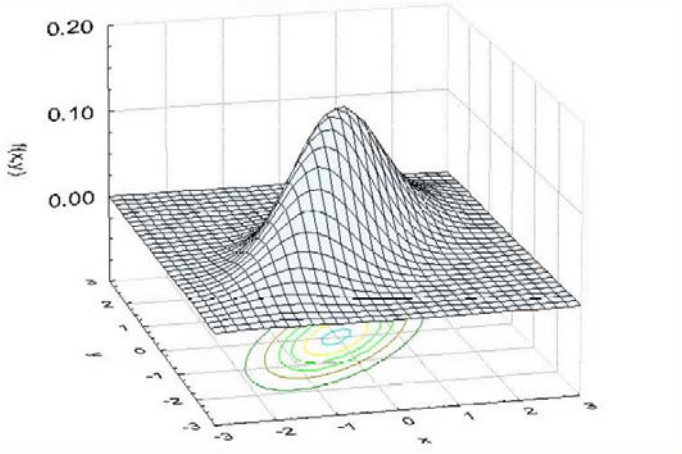
FINITE MIXTURE OF GAUSSIANS

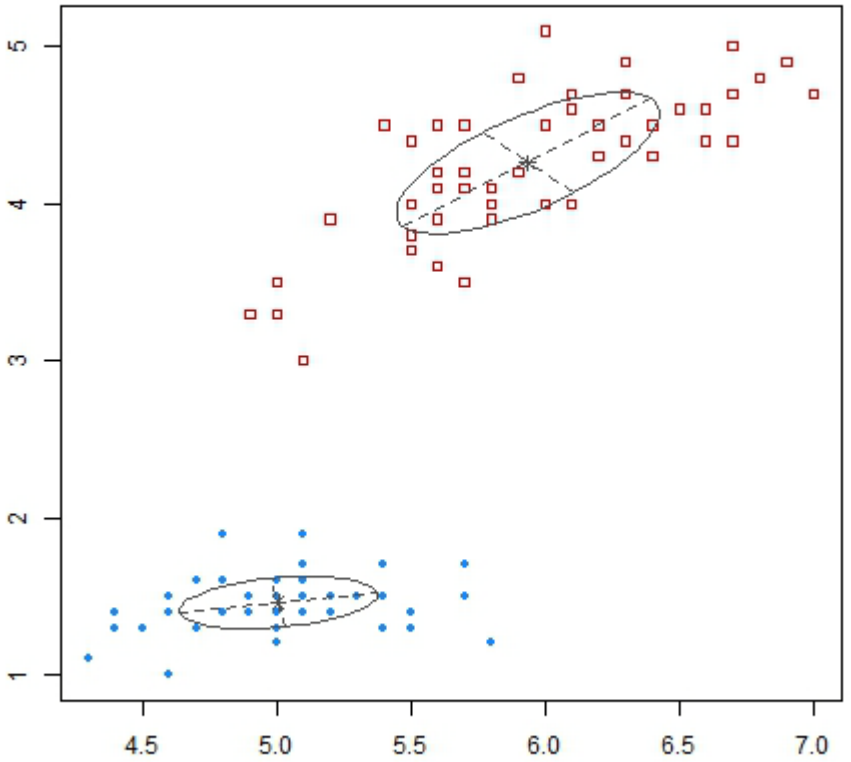
$$f(\boldsymbol{x}) = \sum_{g=1}^{G} p_{\mathrm{g}} \phi(\boldsymbol{x}; \boldsymbol{\mu}_{\mathrm{g}}, \boldsymbol{\Sigma}_{\mathrm{g}})$$

dove:

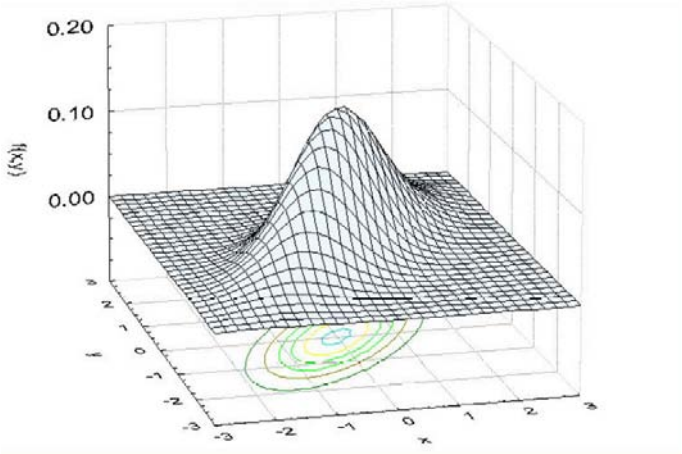
- $\phi(\boldsymbol{x})$ è la funzione di densità normale
- p_{g} sono le probabilità *a priori*
- $\boldsymbol{\mu}_{\mathrm{g}}$ medie delle componenti
- Σ_{g} varianze/covarianze delle componenti

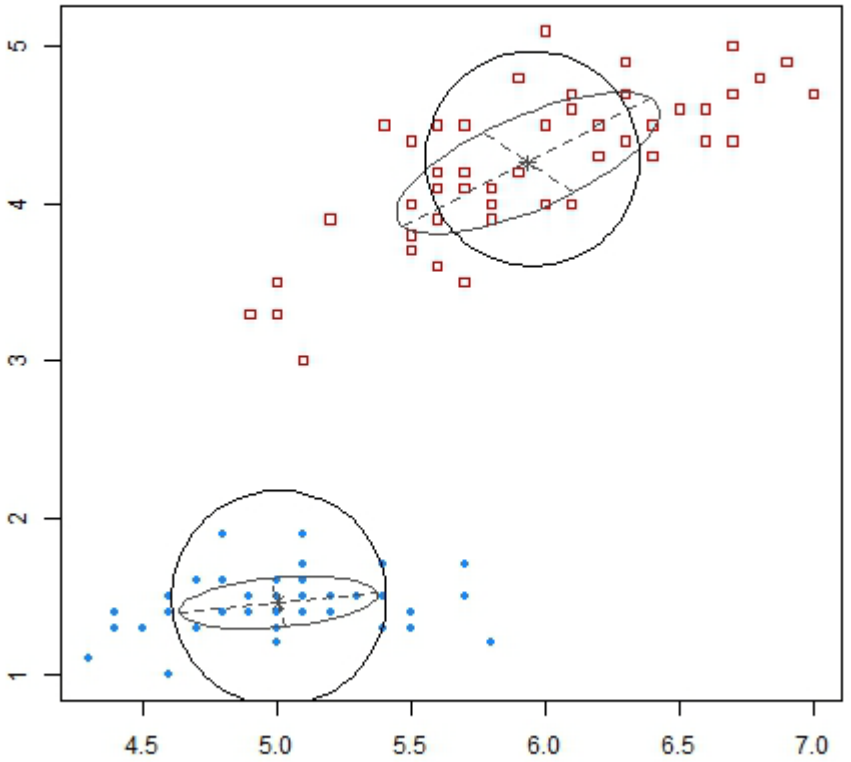
INTUITIVAMENTE …



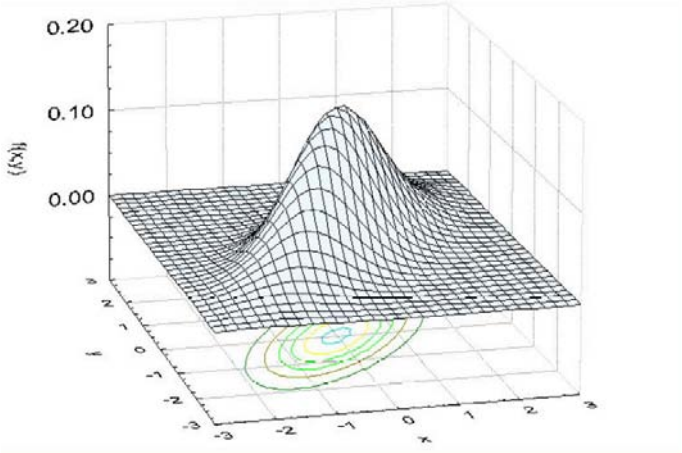


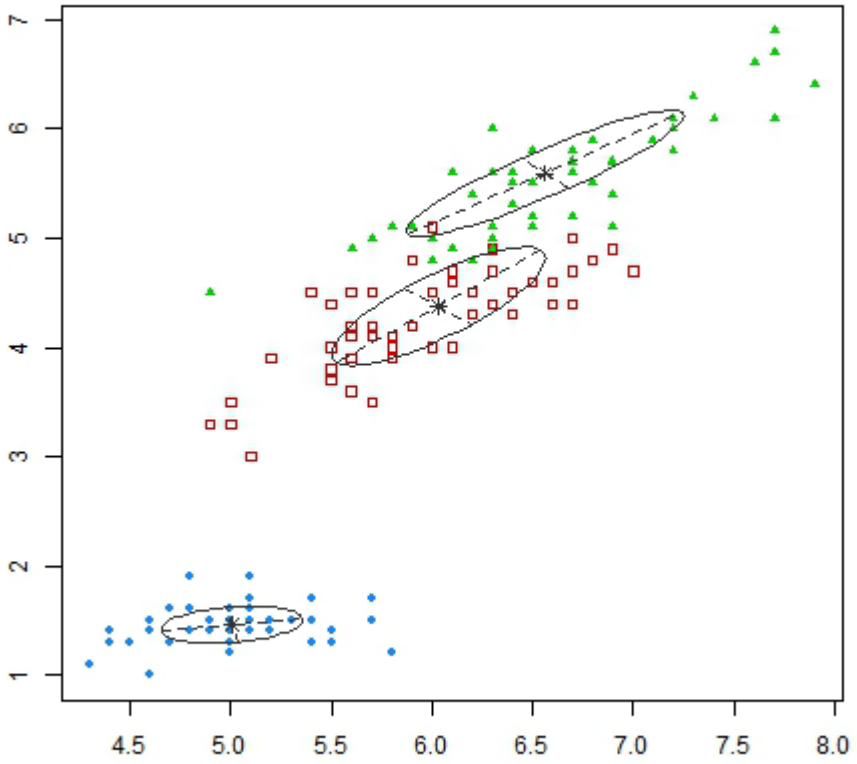
… INTUITIVAMENTE …



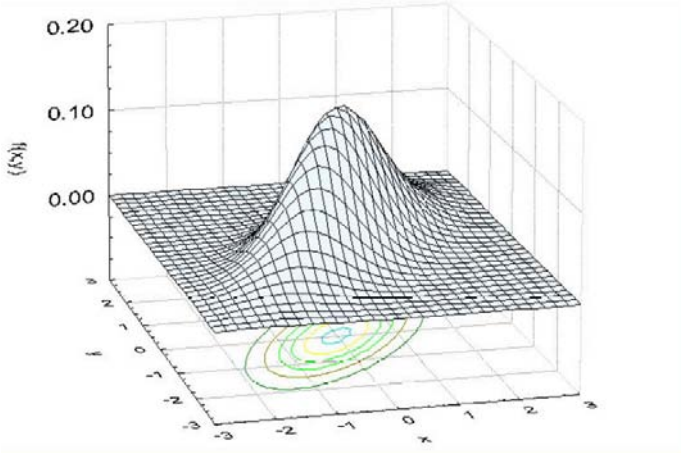


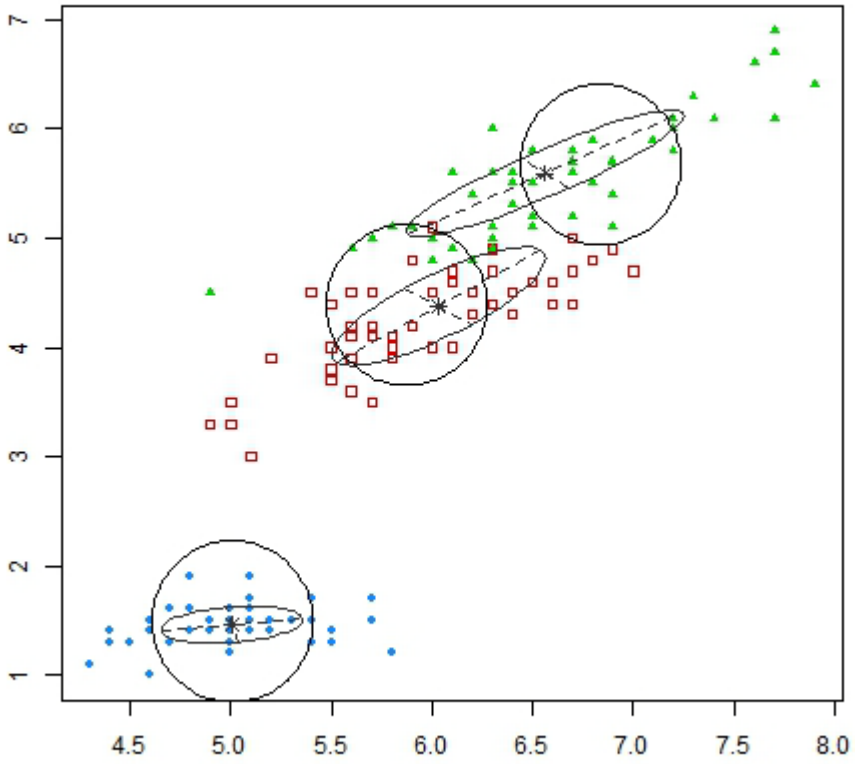
… INTUITIVAMENTE …





… INTUITIVAMENTE …





NOISE E OUTLIER

Per tenere sotto controllo la eventuale presenza di noise e/o oulier può essere aggiunta nella mistura finita una ulteriore componente che rappresenta il noise.

Fraley *et al.* (2012) specificano tale componente attraverso un processo di *Poisson* di primo ordine:

$$f(\boldsymbol{x}) = p_0 \frac{1}{V} + \sum_{g=1}^{G} p_g \phi(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$$

dove:

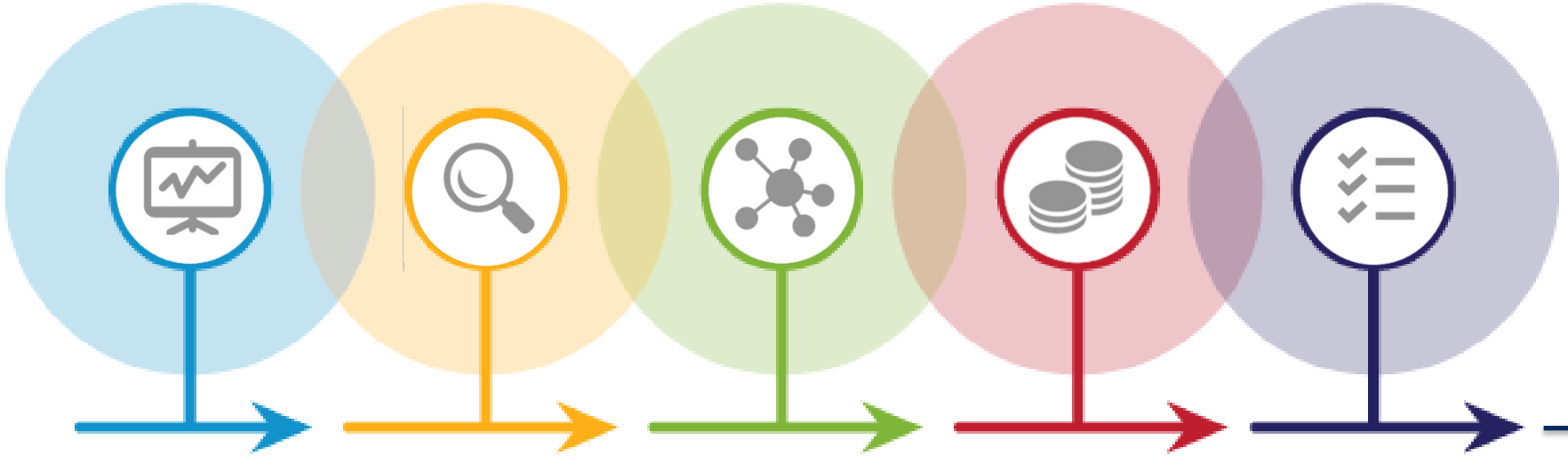
- $\phi(\boldsymbol{x})$ è la funzione di densità normale
- p_g sono le probabilità *a priori*
- V è un threshold tipicamente pari all'ipervolume dei dati

Ambiente di sviluppo

Metodologie a confronto

Obiettivo	Individuazione dei Modelli Organizzativi di Business (MOB) e stima delle probabilità di appartenenza del contribuente ai suddetti	
Ambito	Studi di Settore	Indice Sintetico di Affidabilità
Individuazione cluster (costruzione)	• **Analisi fattoriale (ACP)** • **K-medie sui fattori**	• **Analisi fattoriale (ACP)** • **Misture finite sui fattori**
Stima delle probabilità (applicazione)	• **Analisi discriminante su variabili costituenti i fattori**	
Caratteristica principale	L'assegnazione del contribuente ai gruppi realizzata in costruzione **NON COINCIDE NECESSARIAMENTE** con la probabilità massima di appartenere al medesimo in applicazione	La probabilità di appartenenza ai gruppi realizzata in costruzione **COINCIDE ESATTAMENTE** con quella in applicazione

Il Processo metodologico



1. ANALISI DELLA QUALITÀ DEI DATI

2. INDIVIDUAZIONE DEI FATTORI CARATTERISTICI

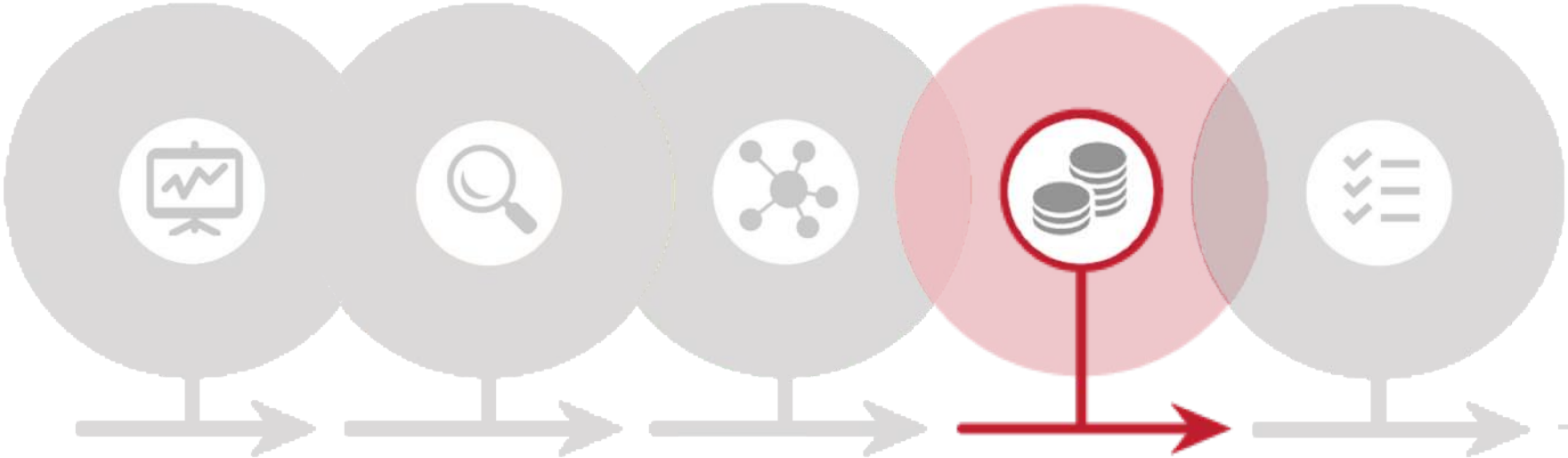
3. INDIVIDUAZIONE DEI GRUPPI OMOGENEI (ANALISI FATTORIALE E CLUSTER ANALYSIS)

4. STIMA DEI RICAVI /COMPENSI, DEL VALORE AGGIUNTO E DEL REDDITO

5. NUOVO SISTEMA DI INDICATORI ELEMENTARI
   - AFFIDABILITÀ
   - ANOMALIE

6. INDICE SINTETICO DI AFFIDABILITÀ

Modelli predittivi



4. STIMA DEI RICAVI /COMPENSI, DEL VALORE AGGIUNTO E DEL REDDITO

6. INDICE SINTETICO DI AFFIDABILITÀ

Considerazioni preliminari

- ✓ *trasformazione logaritmica*
- ✓ *fondamento microeconomico*
- ✓ *introduzione della costante*
- ✓ *da valori assoluti a valori pro-capite*
- ✓ *disegno longitudinale*

Modello di stima

$$y = \alpha + x\beta \ + \varepsilon$$

Specificazione lineare nei logaritmi.

Implicitamente si considera una funzione di produzione Cobb-Douglas:

$$y = \alpha x^{\beta}$$

Uno dei principali vantaggi derivanti da tale specificazione risiede nella relativa interpretazione dei parametri:

- α indica in che misura l'attività produttiva è efficiente, misura la scala di produzione;
- β misura l'elasticità dell'output prodotto rispetto alla corrispondente variabile

Specificazione pro-capite

Si è ritenuto opportuno adottare un modello che faccia riferimento all'input fondamentale utilizzato per la creazione del valore, ovvero il **fattore lavoro**.

In tal senso, il modello proposto prevede che la variabile risposta ed i corrispondenti regressori siano espressi in termini di rapporto rispetto al numero di addetti (i.e. *pro capite*).

I regressori costituiscono degli indicatori che forniscono una immediata lettura dal punto di vista economico.

ANALISI LONGITUDINALE

L’analisi viene condotta a partire dalla banca dati degli studi di settore su un panel in modo da tener conto nella definizione della stima anche del **comportamento individuale** del contribuente nel tempo e dell’**andamento congiunturale**.

Struttura dei dati

Dati di Input

- variabili osservate per impresa/professionista :
  - costo del venduto per addetto
  - spese per lavoro dipendente per addetto
  - valore dei beni strumentali per addetto
  - …
- gruppi omogenei
- misure economiche territoriali (OMI, reddito disponibile, …)
- misure di ciclo economico

Osservazioni ripetute per impresa/professionista

Dati di Output

- Predizione per impresa/professionista
  - ricavi
  - valore aggiunto

Modelli lineari per osservazioni ripetute

Pooled OLS

$$y_{it} = \alpha + x_{it}\beta + \varepsilon_{it}$$

Fixed effects

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}$$

Random effects

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}$$

$$\alpha_i = \alpha + w_i\vartheta + \gamma_i$$

Fixed vs Random

- Nel caso di FE l'inferenza è limitata al comportamento di un particolare insieme di unità.
- Nel caso di modelli ad effetti random, gli stimatori OLS sono ancora non distorti e consistenti, ma non più efficienti.
- Quando è vero il modello ad effetti fissi, gli stimatori OLS non sono più non distorti e consistenti. Tale risultato è conseguenza di un problema di omissioni di variabili dovute al fatto che nello stimatore OLS non sono presenti le dummy per ciascuna unità di analisi.
- La specificazione RE è parsimoniosa, consente di modellizzare l'effetto legato alle singole unità attraverso un solo parametro.
- FE least squares, anche noti come least squares dummy variables (LSDV), prevedono la stima di (n − 1) extra parameteri, pertanto un numero elevato di dummy può aggravre il problema della multicollinearità fra regressori.
- Lo stimatore FE non permette di stimare l'effetto di variabili time-invariant.

Fixed vs Random

Differenza principale

RE assume che la distribuzione degli effetti random non dipende dai regressori mentre FE non richiede tale ipotesi.

In molte applicazioni l'ipotesi di indipendenza può essere violata perchè gli effetti random sono legati alla presenza di variabili time-invariant che sono state omesse dal modello. L'ipotesi di indipendenza è equivalemnte ad essumere che le variabili omesse time-invariant sono indipendenti dai regressori.

È possibile verificare se sia presente una correlazione, ed eventualmente continuare a mantenere una specificazione RE, includendo le medie delle variabili esplicative per gruppo/livello (Mundlak, 1978; Snijders and Bosker, 1999; Hsiao, 2014).

Modello lineare misto

$$\left.\begin{aligned} &\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\ &\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{G}) \\ &\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}) \\ &\boldsymbol{\gamma} \perp \boldsymbol{\varepsilon} \end{aligned}\right\} \quad \boldsymbol{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})$$

dove:

- $\boldsymbol{y}$ il vettore delle osservazioni della variabile risposta;
- $\boldsymbol{\beta}$ vettore degli effetti fissi;
- $\boldsymbol{\gamma}$ vettore degli effetti casuali;
- $\mathbf{X}$ matrice del disegno degli effetti fissi;
- $\mathbf{Z}$ matrice del disegno degli effetti casuali.

Specificazione degli effetti

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_{i}.\boldsymbol{\beta}^{*} + \boldsymbol{w}_{t}\boldsymbol{\eta} + \\ +\delta_1 clu_{it}^{(1)} + \ldots + \delta_{G-1} clu_{it}^{(G-1)} + \gamma_i + \varepsilon_{it}$$

Si è tenuto conto con un'unica funzione delle possibili differenze di risultati economici riconducibili agli aspetti territoriali congiuntamente alle diverse caratteristiche strutturali ed organizzative definite dalla *Cluster Analysis*.

Effetto **casuale** definito dall'*impresa* stessa, analizzando in tal modo il comportamento individuale.

Predizione effetto individuale

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\boldsymbol{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$$

$$\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{G}}\mathbf{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Henderson (1984)

$$\hat{\mathbf{V}} = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}' + \hat{\mathbf{R}}$$

$$\hat{\boldsymbol{G}} = \hat{\sigma}^2_{\mathrm{G}}\boldsymbol{I}_n$$

$$\hat{\boldsymbol{R}} = \hat{\sigma}^2_{\mathrm{R}}\boldsymbol{I}_{nT}$$

Predizione effetto individuale

$$\begin{bmatrix} \widehat{\mathbf{V}}_1 = \mathbf{Z}_1\widehat{\mathbf{G}}\mathbf{Z}_1' + \widehat{\mathbf{R}} & 0 & \cdots & 0 \\ 0 & \ddots & & \\ & & \widehat{\mathbf{V}}_i = \mathbf{Z}_i\widehat{\mathbf{G}}\mathbf{Z}_i' + \widehat{\mathbf{R}} & \\ & & & \ddots \\ 0 & \cdots & 0 & \widehat{\mathbf{V}}_n = \mathbf{Z}_n\widehat{\mathbf{G}}\mathbf{Z}_n' + \widehat{\mathbf{R}} \end{bmatrix}$$

Predizione effetto individuale

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\boldsymbol{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$$

Henderson (1984)

$$\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{G}}\mathbf{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$\hat{\boldsymbol{\gamma}}_{\mathrm{i}} = \hat{\boldsymbol{G}}\mathbf{Z}_{\mathrm{i}}'\,\hat{\boldsymbol{V}}_{\mathrm{i}}^{-1}\left(\boldsymbol{y}_{\mathrm{i}} - \mathbf{X}_{\mathrm{i}}\hat{\boldsymbol{\beta}}\right) \qquad i = 1, \ldots, n$$

Caso particolare T=1

$$\hat{\mathbf{V}}_{\mathrm{i}} = \mathbf{Z}_{\mathrm{i}}\hat{\mathbf{G}}\mathbf{Z}'_{\mathrm{i}} + \hat{\mathbf{R}}_{\mathrm{i}} = \hat{\boldsymbol{\sigma}}_{\mathrm{G}}^{2} + \hat{\boldsymbol{\sigma}}_{\mathrm{R}}^{2}$$

$$\hat{\gamma}_{\mathrm{i}} = \frac{\hat{\boldsymbol{\sigma}}_{\mathrm{G}}^{2}}{\hat{\boldsymbol{\sigma}}_{\mathrm{G}}^{2} + \hat{\boldsymbol{\sigma}}_{\mathrm{R}}^{2}}\,(y_{\mathrm{i}} - \mathbf{x}_{\mathrm{i}}\hat{\boldsymbol{\beta}})$$

Definizione del campione di stima

L'analisi viene condotta a partire dalla banca dati degli studi di settore in relazione ai periodi d'imposta 2007-2014 in modo da tener conto nella definizione della stima anche del comportamento individuale del contribuente nel tempo e dell'andamento congiunturale.

- Panel a partire dalle imprese presenti nell'ultimo periodo di imposta, utilizzato nella fase di costruzione dei gruppi omogenei, ed in maniera **retrospettiva** viene valutata la presenza dei contribuenti fino all'anno di imposta 2007

- Panel «completo» costruito a partire da tutte le imprese presenti nei periodi di imposta osservati

Ambiente di sviluppo

SAS/HPA Procedure Highlights

- **PROC HPLMIXED**
  - High-performance version of PROC MIXED
  - Not to be confused with HPMIXED procedure in SAS/STAT
  - Supports
    - » RANDOM statements
    - » REPEATED statement
    - » Covariance structures from PROC MIXED
  - Sparse MMEQs with > 40,000 unknowns
    - » Impossible in MIXED
    - » 12 hours in HPMIXED
    - » 3 minutes in HPLMIXED

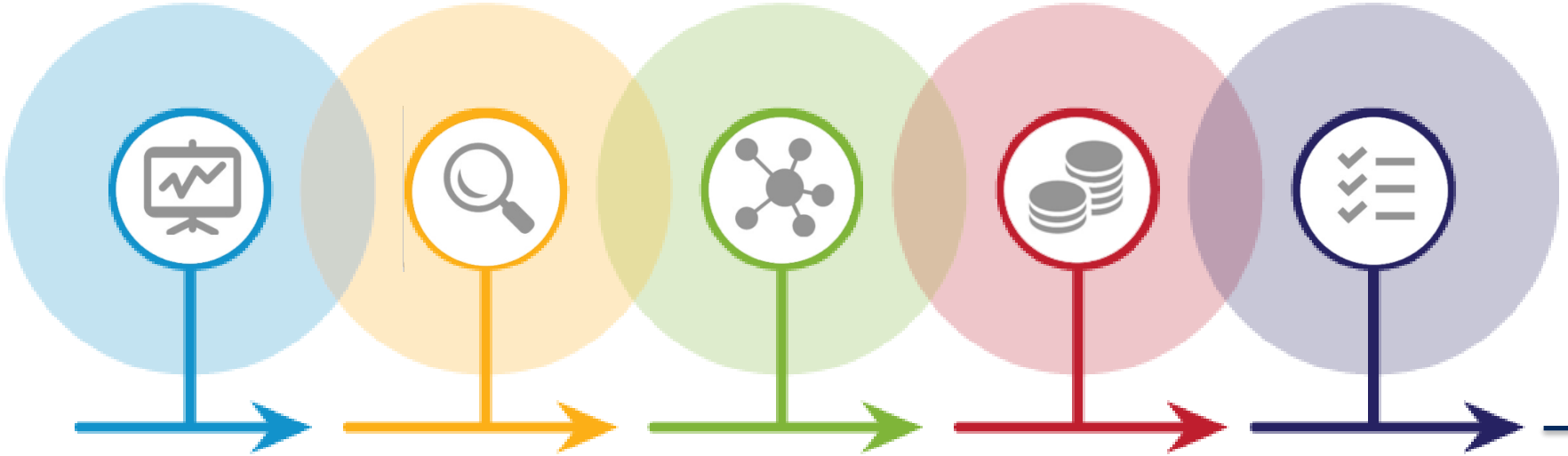
#SASGF11

SAS GLOBAL FORUM

Metodologie a confronto

Obiettivo	Previsione delle basi imponibili in uno specifico periodo di imposta	
Ambito	Studi di Settore	Indice Sintetico di Affidabilità
Variabili obiettivo	Ricavo	Valore aggiunto/Ricavo per addetto (funzione di produzione)
Metodi di stima	• WLS (per gruppo) • indicatori di normalità economica • correttivi crisi	• mixed models (multilevel approach)
Dati	cross-section	panel

Il Processo metodologico



1. ANALISI DELLA QUALITÀ DEI DATI

2. INDIVIDUAZIONE DEI FATTORI CARATTERISTICI

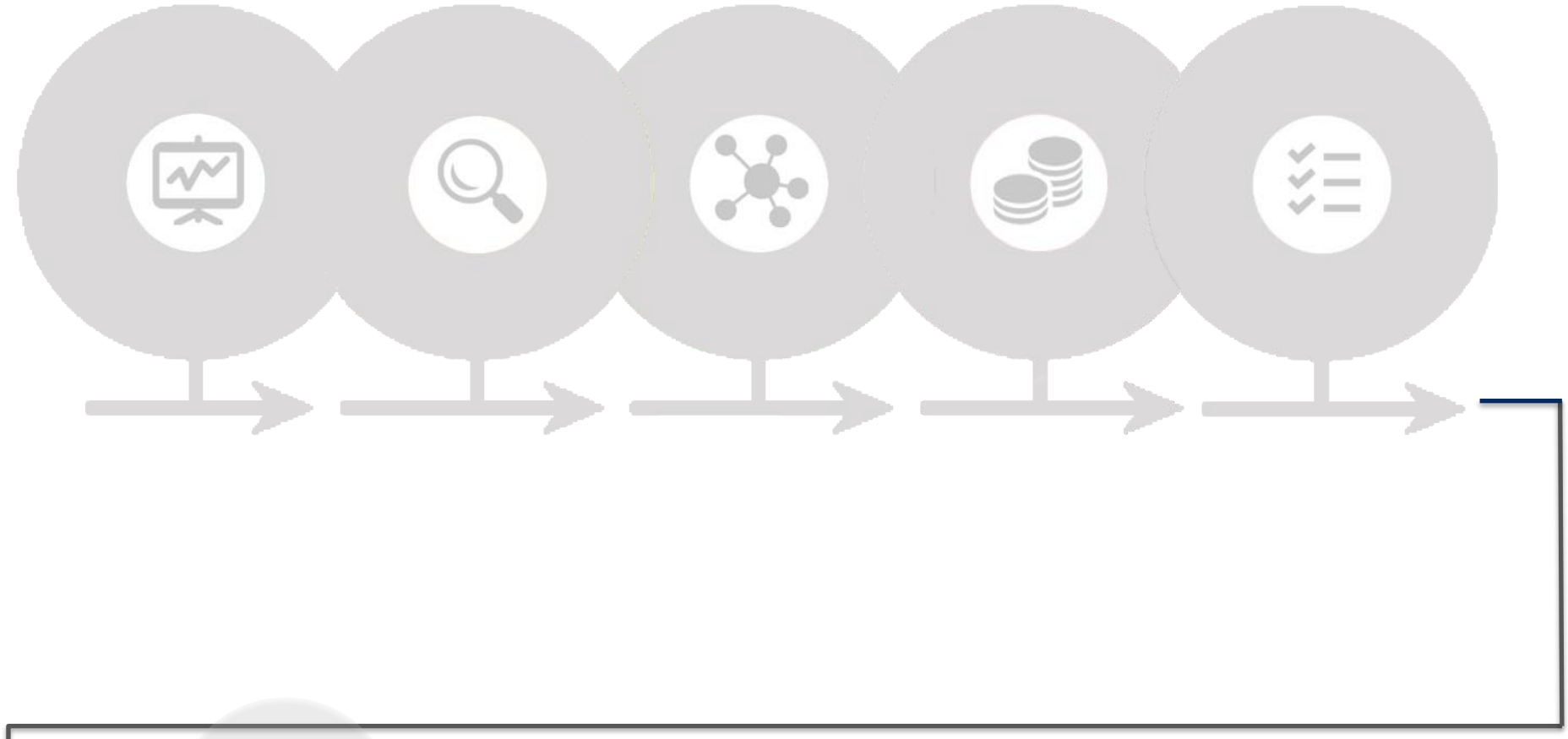
3. INDIVIDUAZIONE DEI GRUPPI OMOGENEI (ANALISI FATTORIALE E CLUSTER ANALYSIS)

4. STIMA DEI RICAVI /COMPENSI, DEL VALORE AGGIUNTO E DEL REDDITO

5. NUOVO SISTEMA DI INDICATORI ELEMENTARI
   - AFFIDABILITÀ
   - ANOMALIE

6. INDICE SINTETICO DI AFFIDABILITÀ

Affidabilità



6. INDICE SINTETICO DI AFFIDABILITÀ

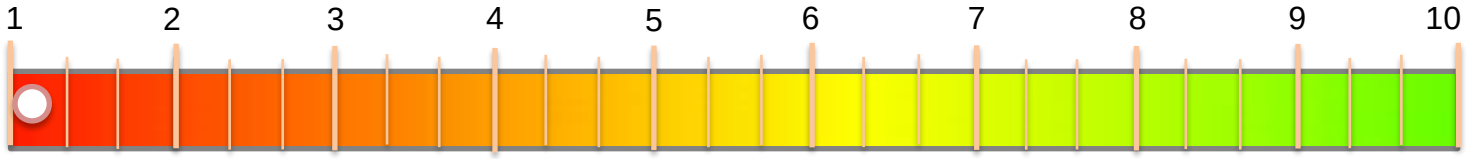
L'INDICE SINTETICO DI AFFIDABILITÀ

RAPPRESENTA IL POSIZIONAMENTO DI OGNI CONTRIBUENTE RISPETTO ALL'AFFIDABILITÀ DEI SUOI COMPORTAMENTI FISCALI.
E' UNA MEDIA SEMPLICE DI INDICATORI ELEMENTARI.

GLI INDICATORI ELEMENTARI PRENDONO IN CONSIDERAZIONE:

- la plausibilità dei ricavi/compensi, del valore aggiunto e del reddito
- l'affidabilità dei dati dichiarati
- le anomalie economiche

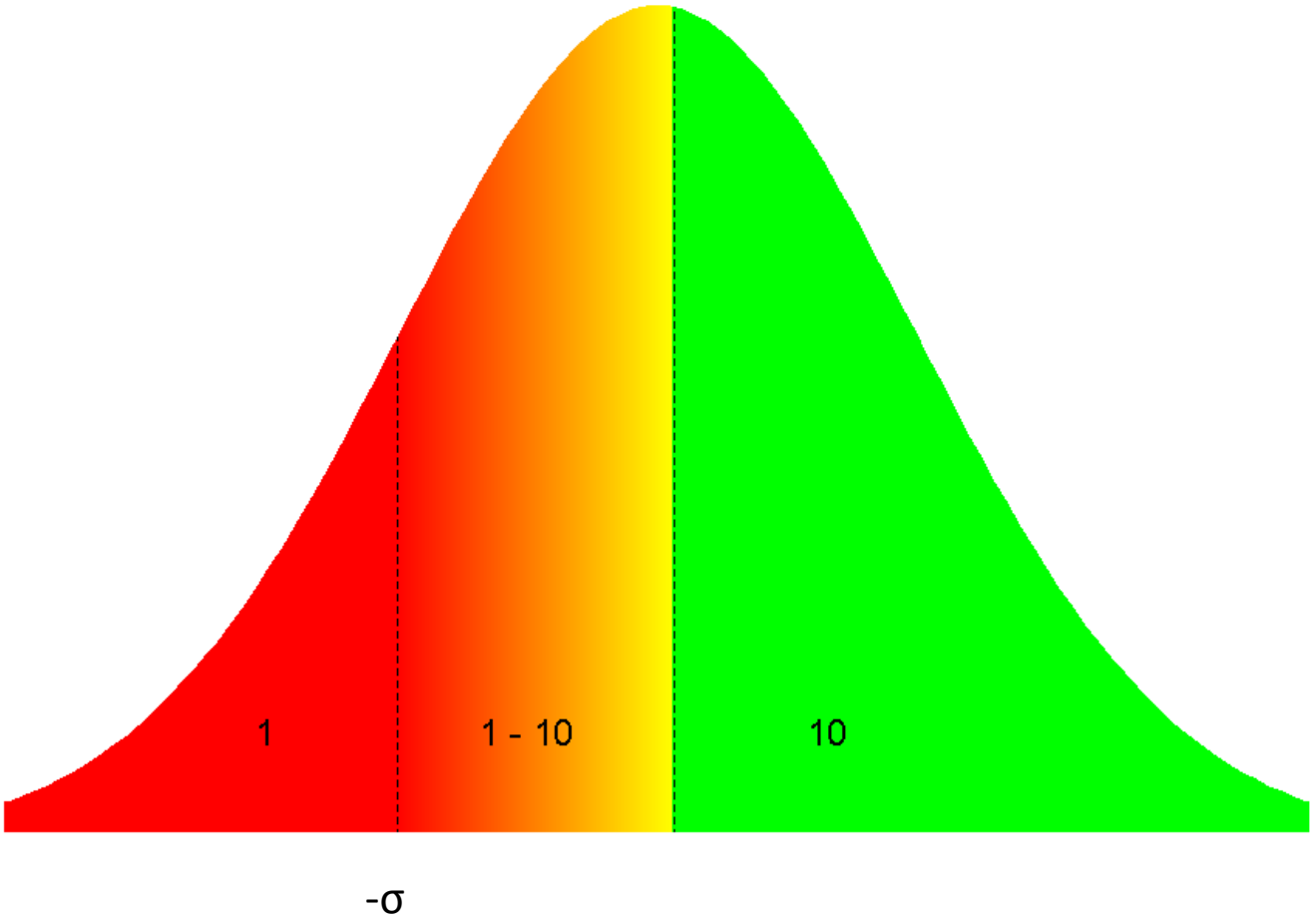
Il valore dell'indice sintetico assegnato al contribuente è compreso tra 1 e 10



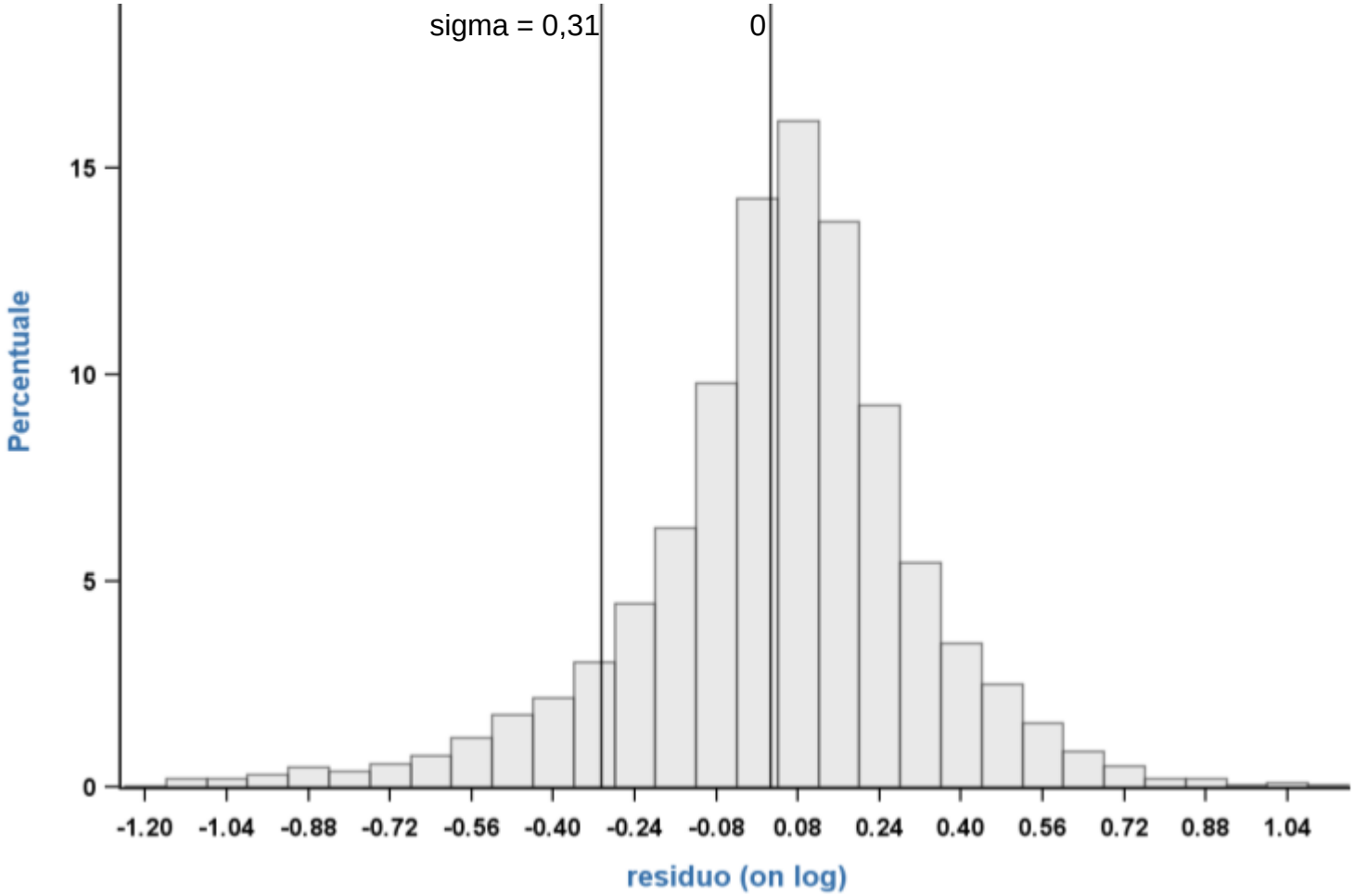
INDICATORI BASATI SU STIME

Calcolo punteggio teorico

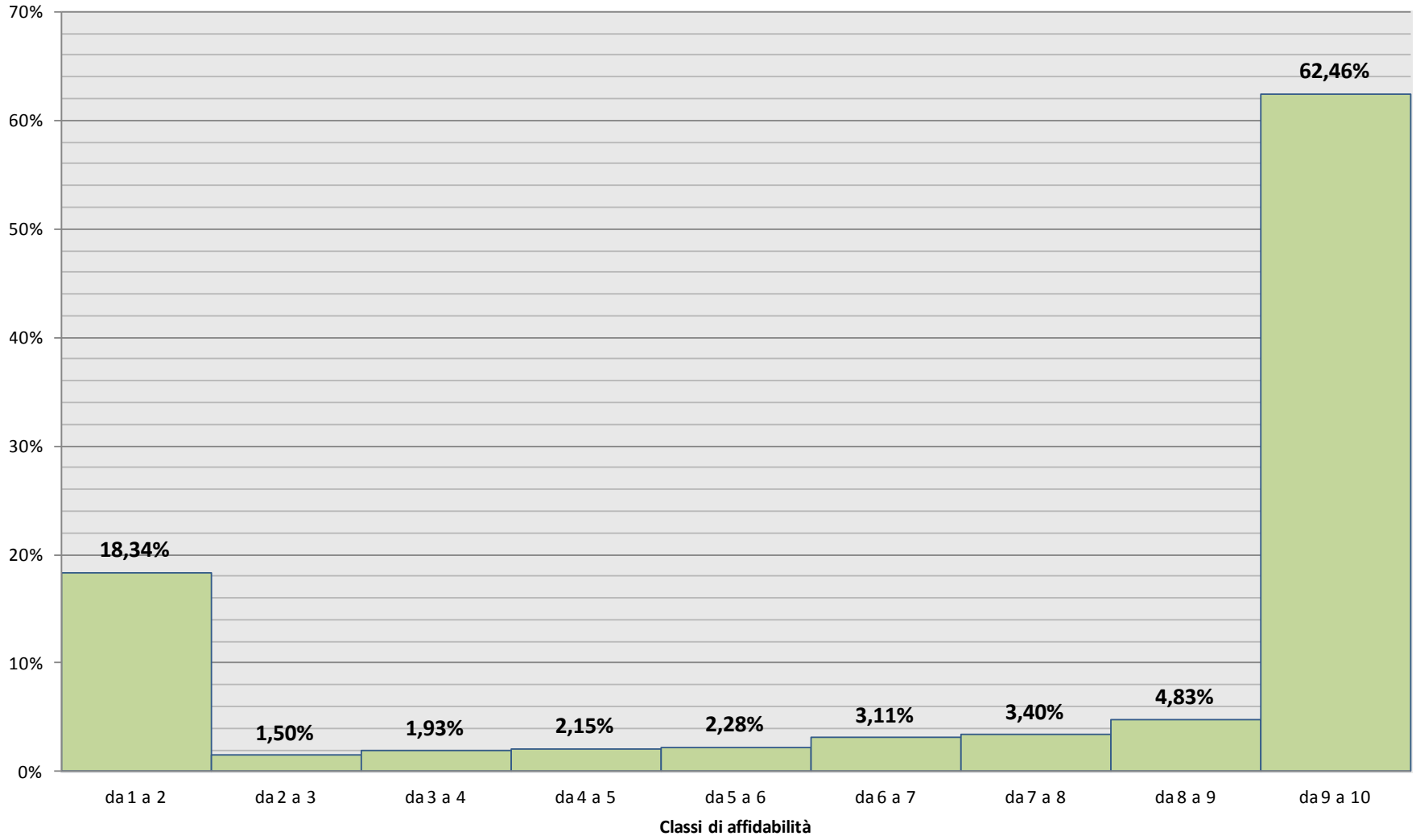
$$\log\left(\frac{VA}{Add}\right) - \log\left(\frac{VA}{Add}\right)^{*}$$



Affidabilità Valore aggiunto per addetto

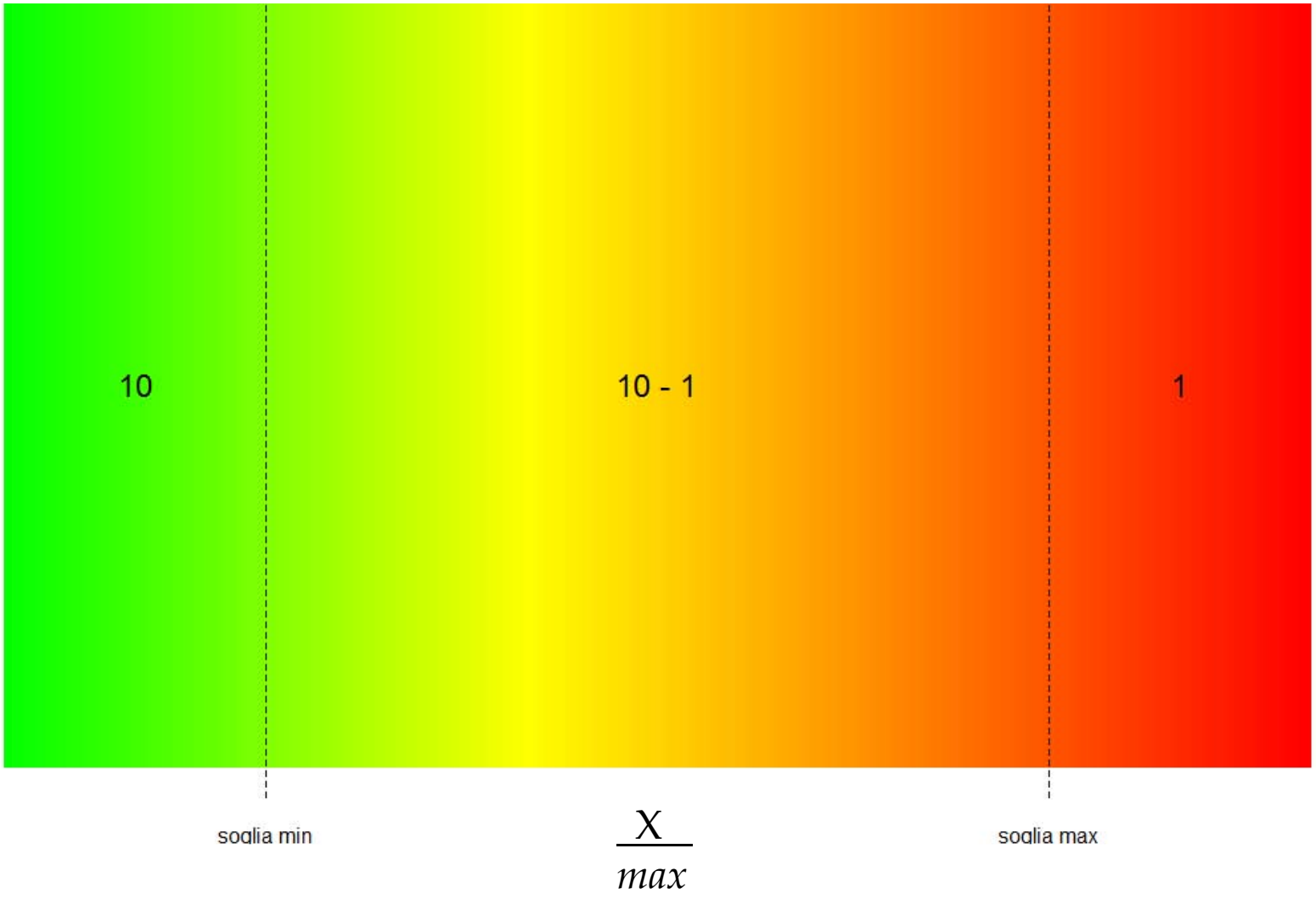


Affidabilità Valore aggiunto per addetto

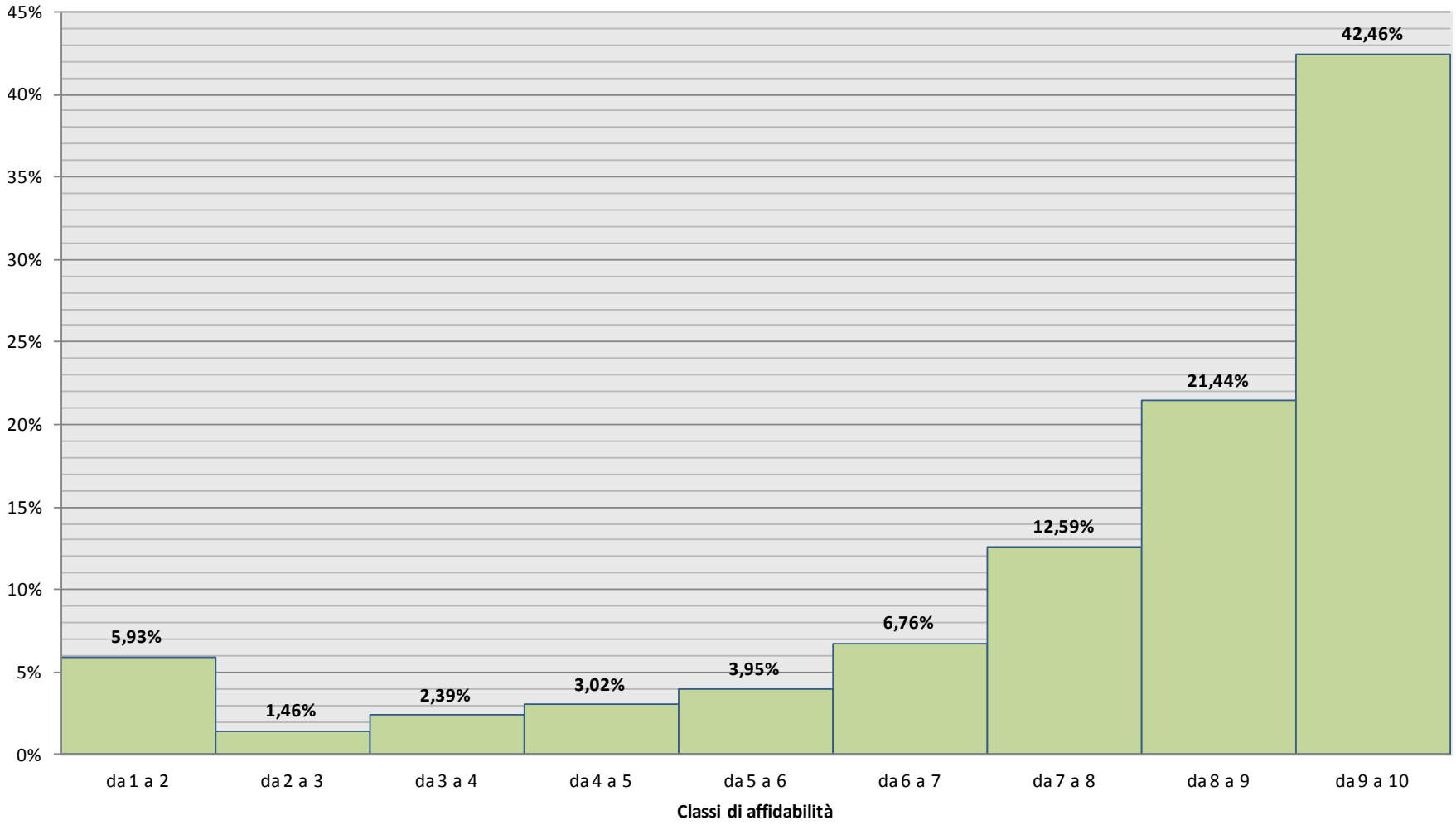


INDICATORI NON BASATI SU STIME

Calcolo punteggio teorico



Affidabilità Durata delle scorte



Giorni	526	468	409	351	292	234	175	117	58	50
Affidabilità	1	2	3	4	5	6	7	8	9	10

GRAZIE PER L'ATTENZIONE!

Arianna Campagna
acampagna@sose.it

Giancarlo Ferrara
gferrara@sose.it