Applicabilità del quality-framework austriaco al SIR italiano

Metodi di Aggregazione degli Indicatori

Romeo Silvestri

1) Introduzione

A partire dai primi anni del ventunesimo secolo, i registri amministrativi hanno assunto una rilevanza sempre maggiore per i fini statistici. In Europa, in particolare, le informazioni contenute nei registri sulle forze lavoro, sulle imposte o in altri strumenti che hanno lo scopo di catalogare la popolazione, sono state integrate nei censimenti nazionali della popolazione residente. Di conseguenza, è sorta la necessità di misurare la correttezza e l'efficacia dei registri amministrativi tramite degli strumenti per la valutazione della qualità. Tali strumenti sono in genere degli indicatori di qualità compositi, volti a cogliere degli aspetti specifici sul processo di costruzione dei registri amministrativi stessi. In questa prospettiva risulta fondamentale la creazione di un forte comparto metodologico legato agli indicatori compositi e di un framework solido per la classificazione e la combinazione dei registri. In questo articolo viene preso in considerazione il caso dello Statistics Austria e dell'ISTAT. Per lo Statistics Austria si prende in esame il quality-framework austriaco presentato nell'Austrian Journal of Statistics del 2016 e si analizzano dei possibili indicatori sintetici che si possono applicare alla prima fase del framework. Per l'ISTAT, invece, si utilizzano le informazioni discusse per lo Statistics Austria e si propone un metodo oggettivo per una eventuale applicabilità del quality-framework austriaco al Sistema Integrato dei Registri (SIR).

2) Quality-Framework austriaco

Per svolgere il censimento della popolazione lo Statistics Austria ha costruito un sistema complesso multifase per quantificare la qualità dei dati. La valutazione è svolta per ciascun registro amministrativo a livello di singola variabile. Il processo di controllo qualità è articolato in 3 fasi progressive:

1) Input: registri amministrativi
2) Process: Database Centrale (CDB)
3) Output: Database Finale (FDB)

La prima fase lavora direttamente con i registri amministrativi che vanno a fornire i dati grezzi di input. Per questa fase lo Statistics Austria ha stabilito che i registri amministrativi possono appartenere a due diversi livelli gerarchici in base alla loro importanza: i registri di base costituiscono la fonte principali dei dati, mentre i registri comparativi servono a convalidare reciprocamente i dati in caso di conflitto tra i registri di base. Nella seconda fase i dati di input vengono aggregati e integrati in macro-gruppi tematici di dati che confluiscono in un unico database centrale (CDB). Infine, nell'ultima fase si va a costituire il database finale FDB contenente le informazioni del CDB e le eventuali imputazioni dei dati mancanti.

In specifiche fasi del framework si vanno ad individuare 4 iperdimensioni di qualità:

- HD^D: iperdimensione della documentazione, racchiude gli aspetti legati alla disponibilità e all'accessibilità dei metadati, inclusa la presenza di documentazione adatta a valutare le fonti che confluiscono nei registri statistici

- HD^P: iperdimensione del pre-processing, riguarda gli errori formali nei dati grezzi e, nello specifico, permette di valutare l'accuratezza dei dati nei registri
- HD^E: iperdimensione delle fonti esterne, presuppone l'esistenza di una o più fonti esterne con cui operare il confronto e permette di valutare la mole di dati che risultano coerenti tra i registri e le fonti
- HD^I: iperdimensione dell'imputazione, è relativa alla qualità del processo di imputazione degli eventuali missing data nell'utlima fase del framework

Ogni iperdimensione coglie un insieme di caratteristiche proprie del processo di costruzione dei registri statistici. Queste servono a realizzare quelli che saranno i veri e propri indicatori di qualità compositi per le 3 fasi del framework. Noi ci focalizzeremo in particolare sull'indicatore di qualità $q_{.j}$ relativo alla prima fase del framework, ma quanto trattato è applicabile anche alle fasi successive.

Nella figura sottostante si riporta una sintesi del quality-framework austriaco con le specifiche iperdimensioni valutate durante le singole fasi:

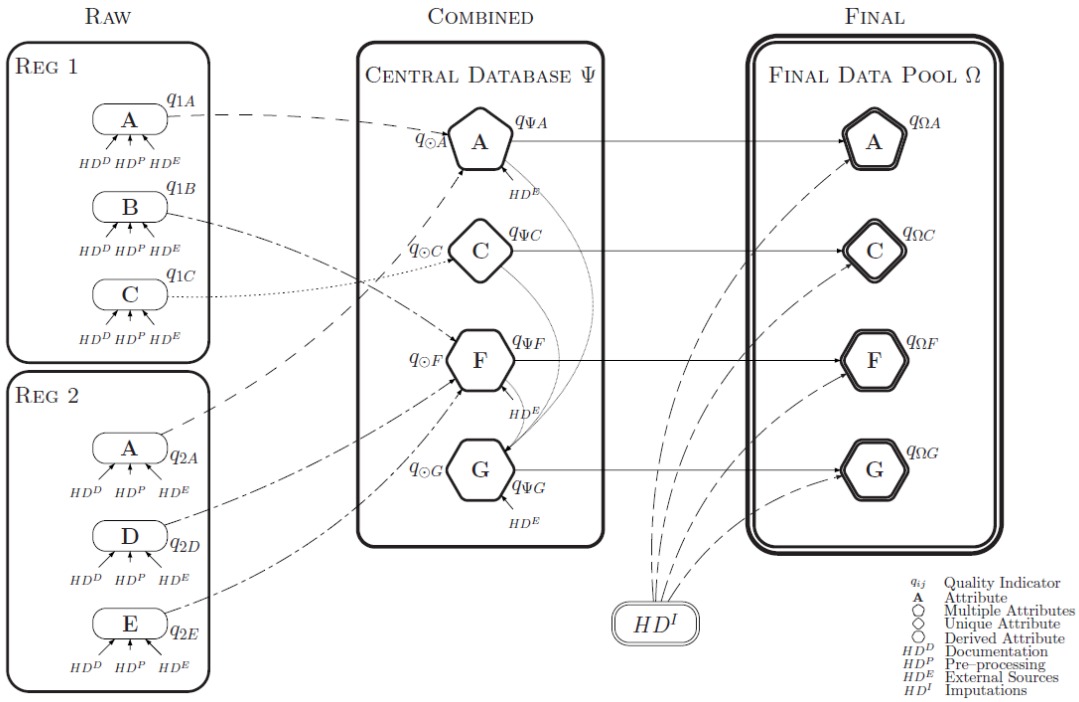

Figure 1: Quality-Framework austriaco

Ogni iperdimensione è formata da numerose sotto-dimensioni rappresentate da indicatori semplici. Tali indicatori vengono combinati per formare uno specifico indicatore composito standardizzato (che assume valori tra 0 e 1) e proprio di una iperdimensione. Questi indicatori compositi vengono a loro volta combinati per costruire degli indicatori ancora più generali (gli indicatori di qualità) che riassumono il livello di qualità per una variabile in una determinata fase del framework.

I metodi tramite cui gli indicatori vengono combinati tra loro costituiscono il tema principale di questa trattazione. Lo Statistics Austria ha infatti proposto un metodo combinatorio semplice basato sulla media aritmetica degli indicatori. Questo metodo non prende in considerazione la diversa rilevanza delle dimensioni e delle iperdimensioni. L'obbiettivo principale è quindi quello di determinare un peso adeguato e oggettivo ad ogni indicatore.

I dati dello Statistics Austria a nostra disposizione fanno riferimento alle dimensioni e iperdimensioni della variabile Stato Civile Legale (LMS). Questi dati sono stati già precedentemente standardizzati tramite numeri indice a base 1, di conseguenza ci occupiamo soltanto della fase di ponderazione degli indicatori.

I valori osservati per gli 11 registri che possiedono le informazioni sul LMS sono riportate nella tabella seguente:

	HD^D									HD^P	HD^E
	SOTTODIMENSIONI HD^D										
REGISTRI	DC	COD	DEF	REL	LB	TIM	AC	TC	DM		
ASR	0	0	1	0	0	1	0,33	0,67	0,33	0,402	0,5
UR	1	1	1	0	1	1	0,67	0	1	0,91	0,981
RPS	0,87	0,87	1	0,62	1	0,8	0,73	0,7	0,63	0,955	0,959
CAR	1	1	1	1	1	1	1	1	0,67	0,973	0,97
CFR	0	0	1	0	0	1	0,33	0,67	0,33	0,898	0,885
CSSR	1	1	1	1	1	1	1	1	0,67	0,454	0,942
CHR	0,67	0,67	1	0,67	0,67	1	0,67	0,78	0,78	0,551	0,887
HPSR	0,35	0,35	1	0,7	0,35	0,81	0,73	0,49	0,59	0,552	0,974
SWR	0,51	0,47	1	0,83	1	0,85	0,81	0,77	0,64	0,783	0,948
CPR	1	1	1	0	1	1	1	1	1	0,67	0,971
TR	1	1	1	1	1	1	1	1	1	0,844	0,91

Figure 2: Iperdimensioni per il LMS

```
# legenda: - registri: Asylum Seekers Register (ASR), Unemployment Register (UR),
#                      Register of Public Servants of the Federal State and the Lander (RPS),
#                      Child Allowance Register (CAR), Central Foreigner Register (CFR),
#                      Central Social Security Register (CSSR), Chambers Register (CHR),
#                      Hospital for Public Servants Register (HPSR),
#                      Register of Social Welfare Recipients (SWR),
#                      Central Population Register (CPR), Tax Register (TR)
#          - sotto-dimensioni: Detect Changes (DC), Cut-off date (COD), Definitions (DEF),
#                              Relevance (REL), Legal Basis (LB), Timeliness (TIM),
#                              Administrative Contr (AC), Technical Contr (TC),
#                              Data Management (DM)
```

Come si può notare dalla tabella, soltanto per l'iperdimensione relativa alla documentazione sono state riportate le sotto-dimensioni, le quali sono indipendenti e facilmente individuabili. Le iperdimensioni associate al pre-processing e alle fonti esterne sono invece formate da sotto-dimensioni aventi la stessa natura e assimilabili in un unico concetto. Per questo motivo si è deciso di trattare le loro sotto-dimensioni come fossero una sola.

Come riportato dallo Statistics Austria si ha infatti che:

$HD^P = \frac{\text{numero di record utilizzabili}}{\text{numero totale di record}}$ e $HD^E = \frac{\text{numero di valori coerenti}}{\text{numero totale di record associati}}$

dove i termini "record utilizzabili" e "valori coerenti" sintetizzano un insieme di sotto-dimensioni che non necessitano un trattamento tra loro differente e, quindi, una diversa ponderazione.

Infine, poiché vogliamo focalizzarci soltanto sulla ponderazione delle dimensioni della prima fase del framework, non considereremo l'iperdimensione per le imputazioni.

Con i seguenti comandi in R si va a creare il dataset di input per l'applicazione dei metodi di ponderazione:

```
# HD_d
# sottodimensioni dell'HD_d
HD_d=c(0,0,1,0,0,1,0.33,0.67,0.33,1,1,1,0,1,1,0.67,0,1,0.87,0.87,1,0.62,1,0.8,0.73,0.7,
       0.63,1,1,1,1,1,1,1,1,0.67,0,0,1,0,0,1,0.33,0.67,0.33,1,1,1,1,1,1,1,1,0.67,0.67,
       0.67,1,0.67,0.67,1,0.67,0.78,0.78,0.35,0.35,1,0.7,0.35,0.81,0.73,0.49,0.59,0.51,
       0.47,1,0.83,1,0.85,0.81,0.77,0.64,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
HD_d=matrix(HD_d,nrow=11,byrow=TRUE)
```

```
colnames(HD_d)=c('DC','COD','DEF','REL','LB','TIM','AC','TC','DM')
rownames(HD_d)=c('ASR','UR','RPS','CAR','CFR','CSSR','CHR','HPSR','SWR','CPR','TR')

# HD_p
HD_p=c(0.402,0.91,0.955,0.973,0.898,0.454,0.551,0.552,0.783,0.67,0.844)

# HD_e
HD_e=c(0.5,0.981,0.959,0.97,0.885,0.942,0.887,0.974,0.948,0.971,0.91)
```

Si vogliono mettere a confronto 4 diversi metodi di ponderazione degli indicatori:

- Metodo Tassonomico di Wroclaw (WTM)
- Indice di Mazziotta-Pareto (MPI)
- Funzione Mean-Min (MMF)
- Analisi delle Componenti Principali (PCA)

Questi metodi sono tra i più utilizzati e discussi in letteratura perché possiedono numerose proprietà auspicabili legate alla teoria degli indicatori. In particolare, i primi tre metodi appartengono alla classe delle funzioni di aggregazione, mentre l'ultimo appartiene all'insieme delle metodologie di sintesi. Nella loro formulazione principale questi metodi prevedono l'uso di uno specifico processo di normalizzazione a monte. In questo caso si è preferito però mantenere per tutti i metodi la standardizzazione con numeri indici a base 1 originaria; questa scelta permette un confronto più immediato tra i diversi metodi e permette di fissare facilmente il dominio dei valori degli indicatori tra 0 e 1.

Andiamo ora a dividere la misurazione dell'indicatore di qualità $q_{.LMS}$ di prima fase in 2 parti. Nella prima parte determiniamo l'iperdimensione della documentazione HD^D_{LMS} tramite le sue sotto-dimensioni, successivamente calcoliamo il valore di $q_{.LMS}$ sfruttando le iperdimensioni HD^D_{LMS}, HD^P_{LMS} e HD^E_{LMS}.

2.1) Determinazione dell'iperdimensione HD^D

2.1.1) Metodo Tassonomico di Wroclaw

Nel metodo di Wroclaw si va a stabilire in primo luogo il vettore dei valori ideali. In questo caso il vettore corrisponde all'insieme dei valori massimi osservati sulle unità per ogni variabile. Conseguentemente si determinano le distanze euclidee D_i di ciascuna unità dall'unità ideale (quella che possiede tutti i valori ideali):

$D_i = \sqrt{\sum_{j=1}^{m}(z_{ij} - z_{0j})^2}$, i=1,...,n distanze euclidee

con z_{ij} , i=1,...,n , j=1,...,m valori standardizzati e $z_{0j} = \max_i\{z_{ij}\}$ valori ideali.

Infine si calcola l'indice sintetico $d_i = \frac{D_i}{D_0}$, dove $D_0 = \overline{D}_0 + 2\sigma_0$, $\overline{D}_0 = \frac{\sum_{i=1}^{n} D_i}{n}$ e $\sigma_0 = \frac{\sum_{i=1}^{n}(D_i - \overline{D}_0)^2}{n}$.

Poiché l'obbiettivo è quello di assegnare come massimo per l'indicatore il valore 1 e come minimo 0, è necessario invertire la polarità dell'indicatore finale. Gli indicatori di Wroclaw risultano così essere pari a $W_i = 1 - d_i$, i=1,...,n.

```
z=HD_d  # valori standardizzati (in questo caso i valori sono già standardizzati)
z0=apply(z,2,max)  # valori ideali

diff_square=(sweep(z,2,z0))^2  # differenze quadratiche con i valori ideali
D=sqrt(apply(diff_square,1,sum))  # distanze euclidee dall'unità ideale
```

```
D0_mean=mean(D)  # media delle distanze
sd_D=sqrt(var(D)*(length(D)-1)/length(D))  # deviazione standard delle distanze
d=D/(D0_mean+2*sd_D)  # indice di Wroclaw

wtm_index=round(1-d,3)  # indice di Wroclaw con polarità inversa
wtm_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.082 0.404 0.705 0.865 0.082 0.865 0.672 0.436 0.635 0.590 1.000
```

2.1.2) Indice di Mazziotta-Pareto

L'indice di Mazziotta-Pareto si basa su una penalizzazione della media aritmetica dei valori standardizzati derivante dal coefficiente di variazione dell'unità in esame. La penalità può essere sia positiva che negativa. Nel caso in questione si può ipotizzare che una maggiore variabilità dei valori delle sotto-dimensioni renda l'indicatore composito più inaffidabile.

Per tale motivo andremo a sottrarre la penalità nel modo seguente:

$MPI_i = \mu_{z_i}(1 - cv_i^2)$, i=1,...,n

con $\mu_{z_i} = \frac{\sum_{j=1}^{m} z_{ij}}{m}$, $cv_i = \frac{\sigma_{z_i}}{\mu_{z_i}}$ e $\sigma_{z_i} = \sqrt{\frac{\sum_{j=1}^{m}(z_{ij}-\mu_{z_i})^2}{m}}$.

```
cv_nc=function(x){  # funzione per il calcolo del campo di variazione
  mean_nc=apply(x,1,mean)
  var_c=apply(x,1,var)
  var_nc=var_c*(dim(x)[2]-1)/(dim(x)[2])
  sd_nc=sqrt(var_nc)
  cv_nc=sd_nc/mean_nc
  return(cv_nc)
}

mean_z=apply(z,1,mean)  # medie dei valori standardizzati
c_z=cv_nc(z)  # cv dei valori standardizzati

mpi_index=mean_z*(1-c_z^2)  # indice di Mazziotta-Pareto
round(mpi_index,3)
```

```
##    ASR     UR    RPS    CAR    CFR   CSSR    CHR   HPSR    SWR    CPR     TR
## -0.061  0.515  0.779  0.952 -0.061  0.952  0.745  0.516  0.722  0.778  1.000
```

Poiché si osservano due registri con valori leggermente negativi, si può pensare di riportare i valori osservati in una scala da 0 a 1. Una possibile soluzione consiste nell'utilizzare degli indici relativi rispetto al campo di variazione del tipo:

$r_i = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$, i=1,...,n

Questa ulteriore normalizzazione degli indicatori può essere fatta in quanto uno dei registri assume già il valore massimo pari a 1. Poiché questo metodo può portare a delle problematiche che tratteremo più avanti, si preferisce portare a 0 i valori dei registri negativi. Questo procedimento si può applicare all'indicatore di Mazziotta-Pareto in quanto l'eccessiva variabilità di un registro può penalizzare la sua performance a tal punto da rendere il registro stesso totalmente inattendibile.

```
for (i in 1:6){
  if (mpi_index[i]<0){
    mpi_index[i]=0
  }
}
mpi_index=round(mpi_index,3)  # mpi normalizzato
mpi_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.000 0.515 0.779 0.952 0.000 0.952 0.745 0.516 0.722 0.778 1.000
```

2.1.3) Funzione Mean-Min

Con il metodo basato sulla funzione Mean-Min l'indicatore ha la seguente forma:

$MMF_i = \mu_{z_i} - \alpha(\sqrt{(\mu_{z_i} - \min_j\{z_{ij}\})^2 + \beta^2} - \beta)$, i=1,...,n

con $0 \leq \alpha \leq 1$ grado di penalizzazione e $\beta \geq 0$ grado di complementarietà degli indicatori.

Per questo motivo risulta cruciale stabilire i valori dei parametri α e β. Non essendoci una regola standard per fissare i parametri, siamo andati a testare varie combinazioni dei parametri di modo che restituiscano dei valori degli indicatori soddisfacenti e in linea con i risultati dei metodi precedenti. In particolare, si ipotizza che gli indicatori delle sotto-dimensioni non siano complementari tra loro ($\beta = 0$) perché indipendenti. Per quanto riguarda il grado di penalizzazione, si è osservato che un valore pari ad $\alpha = 0,5$ restituisce dei risultati bilanciati.

```
min_z=apply(z,1,min)  # minimo delle unità
alpha=0.5
beta=0
mmf_index=mean_z-alpha*((sqrt((mean_z-min_z)^2)+beta^2)-beta)  # indice mmf
mmf_index=round(mmf_index,3)
mmf_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.185 0.371 0.711 0.817 0.185 0.817 0.719 0.473 0.617 0.444 1.000
```

2.1.4) Analisi delle Componenti Principali

Per la costruzione di un indicatore composito esistono numerosi metodi basati sull'analisi delle componenti principali. Il metodo che viene qui trattato utilizza i punteggi (scores) della prima componente principale:

$ACP_i = \sum_{j=1}^{m} a_{j1} z_{ij}$, i=1,...,n

con α_{j1} elementi dell'autovettore associato al più grande autovalore della matrice delle varianze e covarianze degli z_{ij}.

Questo metodo comporta una perdita delle informazioni proporzionale alla varianza non spiegata dalla prima componente principale. Di conseguenza, maggiormente sono correlati tra loro gli indicatori semplici, più efficace risulterà l'indicatore composito.

Svolgiamo quindi l'analisi delle componenti principali prestando attenzione ad escludere dall'analisi le sotto-dimensioni che hanno varianza nulla:

```
library(psych)
pca_HDd=princomp(HD_d[,-3])  # analisi delle componenti principali
summary(pca_HDd)
```

```
## Importance of components:
##                            Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     0.7577513 0.4002492 0.23297830 0.12302354 0.10734438
## Proportion of Variance 0.6978845 0.1947112 0.06597222 0.01839529 0.01400518
## Cumulative Proportion  0.6978845 0.8925957 0.95856793 0.97696322 0.99096840
##                             Comp.6      Comp.7       Comp.8
## Standard deviation     0.070935146 0.048978554 2.992168e-04
## Proportion of Variance 0.006115797 0.002915695 1.088183e-07
## Cumulative Proportion  0.997084196 0.999999891 1.000000e+00
```

La proporzione di varianza spiegata dalla prima componente è pari a 0,6979. Questo significa che l'indicatore composito cattura all'incirca il 70% della variabilità racchiusa negli indicatori semplici.

Andiamo a calcolare i valori degli indicatori. In questo caso la normalizzazione tramite indici relativi rispetto al campo di variazione risulta forzata a causa dei valori restituiti dal metodo basato sulla PCA. A differenza dei risultati dell'indice di Mazziotta-Pareto il valore massimo osservato è però diverso da 1. L'applicabilità di questo metodo di normalizzazione è, d'altra parte, supportata dal fatto che uno dei registri assuma per tutte le dimensioni il valore massimo osservato. E' quindi auspicabile che esso assuma il valore massimo per l'iperdimensione anche nella nuova scala. In questo caso si è costretti però a considerare i registri con performance peggiore come privi di informazione, anche se non si ha un supporto teorico dal punto di vista dell'utilizzo della PCA.

```
pca_index=pca_HDd$scores[,1]
pca_index=(pca_index-min(pca_index))/(max(pca_index)-min(pca_index))
pca_index=round(pca_index,3)
pca_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.000 0.741 0.791 0.965 0.000 0.965 0.649 0.410 0.672 0.845 1.000
```

2.1.5) Confronto tra Indicatori

Si riportano i valori del HD^D_{LMS} per i diversi metodi:

```
cp_table=round(matrix(c(wtm_index,mpi_index,mmf_index,pca_index),nrow=11),3)
colnames(cp_table)=c('WTM','MPI','MMF','PCA')
rownames(cp_table)=c('ASR','UR','RPS','CAR','CFR','CSSR','CHR','HPSR','SWR','CPR','TR')
cp_table
```

```
##        WTM   MPI   MMF   PCA
## ASR  0.082 0.000 0.185 0.000
## UR   0.404 0.515 0.371 0.741
## RPS  0.705 0.779 0.711 0.791
## CAR  0.865 0.952 0.817 0.965
## CFR  0.082 0.000 0.185 0.000
## CSSR 0.865 0.952 0.817 0.965
## CHR  0.672 0.745 0.719 0.649
```

```
## HPSR 0.436 0.516 0.473 0.410
## SWR  0.635 0.722 0.617 0.672
## CPR  0.590 0.778 0.444 0.845
## TR   1.000 1.000 1.000 1.000
```

Analizzando i risultati si osserva che i registri ASR e CFR assumono valori nulli per l'indicatore di Mazziotta-Pareto e per la PCA. Questo fattore è dovuto all'applicazione della post-normalizzazione. Nel caso della normalizzazione per indice relativo rispetto al campo di variazione, si riscontrano una serie di criticità legate alla perdita del confronto con i valori di massimo e minimo teorico. Tra i metodi in esame, il metodo di Wroclaw, l'indice di Mazziotta-Pareto e la funzione Mean-Min con grado di complementarietà $\beta = 0$ sono esenti da questa problematica.

Il metodo con PCA sembra essere il peggiore tra quelli presi in esame. Data la preferenza di mantenere i valori dell'indice di qualità in una scala da 0 a 1, questo metodo necessità sistematicamente di una post-normalizzazione con le conseguenti problematiche sopra descritte. Inoltre, l'utilizzo della sola prima componente principale prevede una necessaria perdita di informazione (in questo caso del 30% in termini di varianza spiegata). Si potrebbe, di conseguenza, provare ad usare un metodo di ponderazione basato sui punteggi fattoriali, ma anche in questo caso si presenterebbe il problema legato alla normalizzazione.

Il metodo basato sulla funzione Mean-Min con questi determinati parametri sembra restituire dei risultati bilanciati e appare particolarmente adatto all'utilizzo con una scala da 0 a 1. Un punto debole di questo metodo è dato tuttavia dalla mancanza di una strategia standard e obbiettiva per determinare il valore dei parametri. Questo fattore allontana il Mean-Min da un approccio totalmente oggettivo.

Prendendo in esame l'indice di Mazziotta-Pareto, il segno della penalizzazione porta ad avere dei valori osservati sbilanciati a sinistra o a destra rispetto alla media aritmetica. Poiché nel nostro caso risulta ideale sottrarre la penalità, tutti i registri avranno sistematicamente dei valori degli indicatori inferiori o uguali a quelli che si otterrebbero con la media. Tale caratteristica può far pensare ad una sottostima degli indicatori di qualità.

Il metodo tassonomico di Wroclaw risulta essere quello più affidabile dal punto di vista applicativo. Una differenza con gli indicatori precedenti consiste nel fatto che il metodo di Wroclaw può potenzialmente assegnare il valore 1 ad un registro anche se questo presenta delle imprecisioni. Questa eventualità si verifica quando un registro assume i valori più elevati rispetto agli altri registri per tutte le sotto-dimensioni o iperdimensioni. Di conseguenza, si potrebbe verificare il caso in cui tutti i registri risultino poco attendibili per una variabile, ma ad uno di questi venga assegnato comunque un indicatore di qualità pari a 1 in seguito alla ponderazione. Avendo come obbiettivo quello di relativizzare i registri, questo fattore non risulta essere un elemento negativo in fase di analisi, bensì un punto di forza.

Fatte queste considerazioni, portiamo avanti l'analisi per tutti i 4 metodi presi in esame, ma consigliamo a livello applicativo l'utilizzo del metodo di Wroclaw.

2.2) Determinazione dell'indicatore di qualità $q_{.j}$

Nella determinazione dell'indicatore di qualità di prima fase utilizziamo gli stessi metodi visti per il calcolo dell'iperdimensione della documentazione. Se è stato quindi precedentemente determinato l'HD^D_{LMS} con uno specifico metodo, quello stesso metodo sarà impiegato per il calcolo del $q_{.LMS}$. Questo modo di procedere è stato scelto per mantenere delle tecniche di analisi coerenti e solide durante le fasi del framework.

2.2.1) Metodo Tassonomico di Wroclaw

Costruiamo il dataset riprendendo l'iperdimensione HD^D trovata con l'indice di Wroclaw e le iperdimensioni HD^P e HD^E:

```
wtm_data=cbind(cp_table[,1],HD_p,HD_e)  # dataset Wroclaw
colnames(wtm_data)[1]='HD_d'
wtm_data
```

```
##       HD_d  HD_p  HD_e
## ASR  0.082 0.402 0.500
## UR   0.404 0.910 0.981
## RPS  0.705 0.955 0.959
## CAR  0.865 0.973 0.970
## CFR  0.082 0.898 0.885
## CSSR 0.865 0.454 0.942
## CHR  0.672 0.551 0.887
## HPSR 0.436 0.552 0.974
## SWR  0.635 0.783 0.948
## CPR  0.590 0.670 0.971
## TR   1.000 0.844 0.910
```

Applichiamo il metodo tassonomico di Wroclaw e mostriamo i risultati:

```
z=wtm_data
z0=apply(z,2,max)

diff_square=(sweep(z,2,z0))^2
D=sqrt(apply(diff_square,1,sum))

D0_mean=mean(D)
sd_D=sqrt(var(D)*(length(D)-1)/length(D))
d=D/(D0_mean+2*sd_D)

wtm_index=round(1-d,3)
wtm_index
```

```
##    ASR     UR    RPS    CAR    CFR   CSSR    CHR   HPSR    SWR    CPR     TR
## -0.034  0.476  0.741  0.882  0.191  0.530  0.526  0.385  0.639  0.554  0.871
```

Si nota che il registro ASR assume un valore negativo. In questo caso una normalizzazione tramite indici relativi rispetto al campo di variazione risulterebbe inopportuna perché assegnerebbe il valore massimo 1 al registro con valore più elevato nonostante non assuma tutti i valori massimi nelle iperdimensioni. L'assegnazione ad un registro del valore massimo per l'indicatore di qualità porterebbe a pensare infatti che tale registro sia esente da criticità o che ne abbia meno rispetto agli altri registri in tutti gli aspetti di analisi.

Un'alternativa possibile consiste nel portare a 0 il valore del registro ASR. Questo procedimento è possibile per le caratteristiche del metodo di Wroclaw. Tale metodo è difatti realizzato in base alle distanze euclidee tra le unità e quella ideale. E' lecito pensare quindi che le unità eccessivamente difformi dall'ideale possano essere escluse dall'analisi. In questo specifico caso, i registri che non rispettano una soglia minima di attendibilità possono non essere presi in considerazione.

Gli indici di Wroclaw finali risultano essere i seguenti:

```
wtm_index[1]=0
wtm_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.000 0.476 0.741 0.882 0.191 0.530 0.526 0.385 0.639 0.554 0.871
```

2.2.2) Indice di Mazziotta-Pareto

Dataset per l'indice di Mazziotta-Pareto:

```
mpi_data=cbind(cp_table[,2],HD_p,HD_e)  # dataset Mazziotta-Pareto
colnames(mpi_data)[1]='HD_d'
mpi_data
```

```
##       HD_d  HD_p  HD_e
## ASR  0.000 0.402 0.500
## UR   0.515 0.910 0.981
## RPS  0.779 0.955 0.959
## CAR  0.952 0.973 0.970
## CFR  0.000 0.898 0.885
## CSSR 0.952 0.454 0.942
## CHR  0.745 0.551 0.887
## HPSR 0.516 0.552 0.974
## SWR  0.722 0.783 0.948
## CPR  0.778 0.670 0.971
## TR   1.000 0.844 0.910
```

Applichiamo il metodo per il calcolo dell'indice:

```
z=mpi_data
mean_z=apply(z,1,mean)
c_z=cv_nc(z)

mpi_index=mean_z*(1-c_z^2)
mpi_index=round(mpi_index,3)
mpi_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.145 0.750 0.890 0.965 0.297 0.714 0.702 0.617 0.807 0.787 0.914
```

2.2.3) Funzione Mean-Min

Carichiamo il dataset e applichiamo il metodo basato sulla funzione Mean-Min:

```
mmf_data=cbind(cp_table[,3],HD_p,HD_e)  # dataset Mean-Min
colnames(mmf_data)[1]='HD_d'
mmf_data
```

```
##       HD_d  HD_p  HD_e
## ASR  0.185 0.402 0.500
## UR   0.371 0.910 0.981
## RPS  0.711 0.955 0.959
## CAR  0.817 0.973 0.970
## CFR  0.185 0.898 0.885
## CSSR 0.817 0.454 0.942
## CHR  0.719 0.551 0.887
## HPSR 0.473 0.552 0.974
```

```
## SWR  0.617 0.783 0.948
## CPR  0.444 0.670 0.971
## TR   1.000 0.844 0.910
```

```
z=mmf_data
mean_z=apply(z,1,mean)
min_z=apply(z,1,min)
alpha=0.5
beta=0

mmf_index=mean_z-alpha*((sqrt((mean_z-min_z)^2)+beta^2)-beta)
mmf_index=round(mmf_index,3)
mmf_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.274 0.562 0.793 0.868 0.420 0.596 0.635 0.570 0.700 0.570 0.881
```

2.2.4) Analisi delle Componenti Principali

Mostriamo i risultati per il metodo basato sulle componenti principali:

```
pca_data=cbind(cp_table[,4],HD_p,HD_e)  # dataset Componenti Principali
colnames(pca_data)[1]='HD_d'
pca_data
```

```
##       HD_d  HD_p  HD_e
## ASR  0.000 0.402 0.500
## UR   0.741 0.910 0.981
## RPS  0.791 0.955 0.959
## CAR  0.965 0.973 0.970
## CFR  0.000 0.898 0.885
## CSSR 0.965 0.454 0.942
## CHR  0.649 0.551 0.887
## HPSR 0.410 0.552 0.974
## SWR  0.672 0.783 0.948
## CPR  0.845 0.670 0.971
## TR   1.000 0.844 0.910
```

```
pca_qi=princomp(pca_data)
summary(pca_qi)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3
## Standard deviation     0.3614315 0.1911529 0.08377425
## Proportion of Variance 0.7499427 0.2097673 0.04029000
## Cumulative Proportion  0.7499427 0.9597100 1.00000000
```

```
# la prima componente possiede all'incirca il 75% di varianza spiegata
```

```
pca_index=pca_qi$scores[,1]
pca_index=round(pca_index,3)
pca_index
```

```
##    ASR     UR    RPS    CAR    CFR   CSSR    CHR   HPSR    SWR    CPR     TR
## -0.783  0.160  0.211  0.381 -0.560  0.249 -0.038 -0.237  0.056  0.196  0.366
```

Con l'utilizzo di questo metodo sorge in questo caso un problema legato alla standardizzazione dell'indicatore in una scala da 0 a 1. Per il motivo discusso nel paragrafo 2.2.1 con il metodo di Wroclaw, non risulta ottimale normalizzare l'indicatore tramite indice relativo rispetto al campo di variazione. Essendo la prima componente centrata in 0 non si può nemmeno optare per la stessa soluzione utilizzata col metodo di Wroclaw. Si osserva che questo problema si potrebbe verificare anche per l'indice basato sulla funzione Mean-Min per $\beta > 0$.

Decidiamo comunque di applicare una normalizzazione tramite indice relativo rispetto al campo di variazione per potere successivamente confrontare l'indicatore trovato con gli altri metodi:

```
pca_index=(pca_index-min(pca_index))/(max(pca_index)-min(pca_index))
pca_index=round(pca_index,3)
pca_index
```

```
##   ASR    UR   RPS   CAR   CFR  CSSR   CHR  HPSR   SWR   CPR    TR
## 0.000 0.810 0.854 1.000 0.192 0.887 0.640 0.469 0.721 0.841 0.987
```

2.2.5) Confronto tra Indicatori

```
cp_table=round(matrix(c(wtm_index,mpi_index,mmf_index,pca_index),nrow=11),3)
colnames(cp_table)=c('WTM','MPI','MMF','PCA')
rownames(cp_table)=c('ASR','UR','RPS','CAR','CFR','CSSR','CHR','HPSR','SWR','CPR','TR')
cp_table
```

```
##        WTM   MPI   MMF   PCA
## ASR  0.000 0.145 0.274 0.000
## UR   0.476 0.750 0.562 0.810
## RPS  0.741 0.890 0.793 0.854
## CAR  0.882 0.965 0.868 1.000
## CFR  0.191 0.297 0.420 0.192
## CSSR 0.530 0.714 0.596 0.887
## CHR  0.526 0.702 0.635 0.640
## HPSR 0.385 0.617 0.570 0.469
## SWR  0.639 0.807 0.700 0.721
## CPR  0.554 0.787 0.570 0.841
## TR   0.871 0.914 0.881 0.987
```

3) Quality-Framework applicato al SIR

Si vuole valutare una possibile applicazione del quality-framework austriaco al Registro Base degli Individui delle famiglie e delle convivenze (RBI), ovvero uno dei registri statistici di base del SIR italiano. Per i dettagli relativi ai registri utilizzati e al comparto tecnico si rimanda all'articolo riportato in bibliografia

sull'applicabilità del framework austriaco. Nel seguito si fa riferimento a tale articolo e si propone un metodo di aggregazione oggettivo degli indicatori rispetto a quello ipotizzato e in linea con quanto trattato per il caso austriaco.

Come analizzato nell'articolo sopra citato, non si ha a disposizione nell'immediato delle fonti esterne al SIR da mettere a confronto. Non risulta quindi possibile valutare l'iperdimensione HD^E; d'altra parte è stata proposta una possibile soluzione che prevede l'utilizzo del Master Sample (MS). Si procede allora con lo studio basato sull'iperdimensione HD^D e HD^P.

Per il caso italiano si prendono in considerazione le osservazioni su 6 registri per la variabile Sesso. Per questa variabile si hanno a disposizione i valori grezzi delle sotto-dimensioni dell'HD^D facenti riferimento alle definizioni prodotte dall'Eurostat nel 2020.

Mostriamo la tabella con le sotto-dimensioni non normalizzate dell'HD^D e con i valori normalizzati per l'HD^P:

	HD^D					HD^P
	SOTTODIMENSIONI HD^D					
REGISTRI	CON_1	CON_2	CHI	PUN	TEM	
LAC	1	1	3	-4	99	0,99999997
AT	1	1	3	0	99	1
MAEAIE	1	1	3	-19	195	1
INPSDMAG	1	1	3	-8	350	0,99964194
INPSRAPPLAV	1	1	3	-12	163	0,99988127
ISCRNAS	1	1	3	0	329	0,954

Figure 3: Iperdimensioni per la variabile Sesso

Riportiamo i corrispondenti comandi in R per la creazione del dataset:

```
#sottodimensioni dell'HD_d
HD_d=c(1,1,3,-4,99,1,1,3,0,99,1,1,3,-19,195,1,1,3,-8,350,1,1,3,-12,163,1,1,3,0,329)
HD_d=matrix(HD_d,nrow=6,byrow=TRUE)

colnames(HD_d)=c('CON_1','CON_2','CHI','PUN','TEM')
rownames(HD_d)=c('LAC','AT','MAEAIE','INPSDMAG','INPSRAPPLAV','ISCRNAS')

#HD_p
HD_p=c(0.99999997,1,1,0.99964194,0.99988127,0.954)
```

```
# legenda: - registri: Liste Anagrafiche Comunali (LAC),
#                      Agenzia delle Entrate - Anagrafe delle Persone Fisiche (AT),
#                      Ministero Affari Esteri - Archivio Italiani all'Estero (MAEAIE),
#                      INPS - DMAG dichiarazione sulla manodopera agricola (INPSDMAG),
#                      INPS - Rapporti di lavoro domestico (INPSRAPPLAV),
#                      Iscritti all'Anagrafe per nascita (ISCRNAS)
#          - sotto-dimensioni: Confrontabilità di tipo 1 - Interruzioni nella
#                              serie temporale (CON_1), Confrontabilità di tipo 2 -
#                              disponibilità dello storico dei tracciati record (CON_2),
#                              Chiarezza (CHI), Puntualità (PUN), Tempestività (TEM)
```

Si vogliono normalizzare i valori tramite numeri indici in base 1 rispetto al massimo teorico nel modo seguente:

$z_{ij} = \frac{x_{ij}}{x_{0j}}$, i=1,...,n , j=1,...,m con x_{0j} massimi teorici.

Questo metodo permette di svincolare gli indicatori dall'unità di misura e porre una scala da 0 a 1 per tutte le sotto-dimensioni. In particolare, i valori di massimo teorico si basano sul metodo di misurazione utilizzato per i valori grezzi. Poniamo un valore di massimo pari a 1 per la confrontabilità di tipo 1 essendo una variabile dicotomica. Per la confrontabilità di tipo 2 e per la chiarezza si fissa un massimo in base al numero delle modalità per ciascuna delle due variabili ordinali rappresentate. Per quanto riguarda la puntualità e la tempestività, queste non hanno propriamente dei valori di massimo teorico. Per entrambe le sotto-dimensioni abbiamo stabilito come massimo teorico 365 giorni poiché un valore maggiore porterebbe a rendere un ipotetico registro inutilizzabile al fine della determinazione di una variabile. Puntualità e tempestività misurano infatti il numero di giorni necessari per rendere rispettivamente disponibile e calcolabile la variabile. Un eventuale ritardo superiore ad un anno porterebbe così all'impossibilità del confronto tra i registri. Per queste due sotto-dimensioni si inverte, inoltre, il segno della polarità in quanto è auspicabile un valore che risulti il più basso possibile.

Mostriamo la tabella dei dati normalizzati relativi alle sotto-dimensioni dell'HD^D per la variabile Sesso:

```
HD_d_max=c(1,2,5,365,365)
HD_d_np=t(t(HD_d)/HD_d_max)

HD_d_norm=cbind(HD_d_np[,1:3],1-HD_d_np[,4:5])
HD_d_norm
```

```
##             CON_1 CON_2 CHI      PUN        TEM
## LAC             1   0.5 0.6 1.010959 0.72876712
## AT              1   0.5 0.6 1.000000 0.72876712
## MAEAIE          1   0.5 0.6 1.052055 0.46575342
## INPSDMAG        1   0.5 0.6 1.021918 0.04109589
## INPSRAPPLAV     1   0.5 0.6 1.032877 0.55342466
## ISCRNAS         1   0.5 0.6 1.000000 0.09863014
```

Si nota che la sotto-dimensione della puntualità assume dei termini fuori dalla scala teorica. Numerosi registri rendono per l'appunto disponibile in anticipo i dati per la variabile Sesso. Poiché questo è un elemento che attribuisce ulteriore qualità al registro, si decide di lasciare i valori sulla puntualità fuori scala.

Successivamente applichiamo un metodo di ponderazione delle sotto-dimensioni per determinare l'iperdimensione relativa alla documentazione. Avendo restituito i risultati migliori durante l'applicazione del quality-framework austriaco, si decide di utilizzare il metodo tassonomico di Wroclaw:

```
z=HD_d_norm
z0=apply(z,2,max)

diff_square=(sweep(z,2,z0))^2
D=sqrt(apply(diff_square,1,sum))

D0_mean=mean(D)
sd_D=sqrt(var(D)*(length(D)-1)/length(D))
d=D/(D0_mean+2*sd_D)

wtm_index=round(1-d,3)
wtm_index
```

```
##         LAC          AT      MAEAIE    INPSDMAG INPSRAPPLAV     ISCRNAS
##       0.950       0.937       0.683       0.170       0.787       0.237
```

Visualizziamo il dataset finale con le iperdimensioni e calcoliamo l'indicatore di qualità di prima fase:

```
wtm_data=cbind(wtm_index,HD_p)
colnames(wtm_data)[1]='HD_d'
wtm_data
```

```
##              HD_d      HD_p
## LAC         0.950 1.0000000
## AT          0.937 1.0000000
## MAEAIE      0.683 1.0000000
## INPSDMAG    0.170 0.9996419
## INPSRAPPLAV 0.787 0.9998813
## ISCRNAS     0.237 0.9540000
```

```
z=wtm_data
z0=apply(z,2,max)

diff_square=(sweep(z,2,z0))^2
D=sqrt(apply(diff_square,1,sum))

D0_mean=mean(D)
sd_D=sqrt(var(D)*(length(D)-1)/length(D))
d=D/(D0_mean+2*sd_D)

wtm_index=round(1-d,3)
wtm_index
```

```
##         LAC          AT      MAEAIE    INPSDMAG INPSRAPPLAV     ISCRNAS
##       1.000       0.986       0.719       0.180       0.829       0.249
```

Utilizzando il metodo di Wroclaw si può vedere che l'iperdimensione HD^D è quella più influente per la determinazione dell'indicatore di qualità. L'iperdimensione della documentazione ha infatti una maggiore variabilità rispetto all'iperdimensione del pre-processing e questo fa in modo che abbia un impatto maggiore.

Il LAC risulta essere il registro con l'indicatore di qualità più elevato e pari al massimo sulla scala di misura di riferimento. Il LAC si pone quindi come standard di qualità in prima fase per la variabile Sesso. Questo fenomeno si registra quando, come in questo caso, un registro assume il valore di massimo osservato per tutte le dimensioni.

Indicatori di qualità elevati ($> 0{,}7$) si registrano anche per i registri AT, INPSRAPPLAV e MAEAIE. Poco attendibili per la variabile Sesso sono invece i registri INPSDMAG e ISCRNAS.

4) Conclusioni

Abbiamo analizzato alcune delle possibili tecniche di normalizzazione e ponderazione per costruire degli indicatori compositi volti a valutare la qualità dei registri amministrativi. Questi metodi sono stati testati inizialmente sul quality-framework austriaco e rielaborati per l'applicazione sul SIR italiano. Si è visto che ogni metodo di ponderazione presenta dei vantaggi e degli svantaggi, ma ognuno di essi è preferibile per delle determinate situazioni. Nel caso in esame, dovendo porre gli indicatori su una scala da 0 a 1, abbiamo optato per il metodo tassonomico di Wroclaw unito ad una pre-normalizzazione tramite numeri indici.

Per chi affronterà successivamente questo tema, si suggerisce di andare a testare ulteriori metodi di ponderazione come l'indice di Jevons o l'analisi multicriteria. Si potrebbe inoltre pensare ad un diverso metodo

di standardizzazione che non preveda necessariamente la scala dei valori da noi utilizzata. Questo aspetto permetterbbe di evitare la post-normalizzazione degli indicatori.

Un'ulteriore analisi per la scelta del metodo da impiegare potrebbe essere portata avanti con un'analisi di sensitività. Si riporta in bibliografia il libro di riferimento dell'OECD che mostra una serie di esempi applicativi su questo tema.

Bibliografia

Eva-Maria Asamer, Franz Astleithner, Henrik Rechta, Manuela Lenk, Mathias Moser, Predrag Ćetković, Stefan Humer, 2016. Quality Assessment for Register-based Statistics-Results for the Austrian Census 2011 (Austrian Journal of Statistics)

Eurostat (2020). European Statistical System handbook for quality and metadata reports

Filomena Maggino, 2017. Complexity in Society: From Indicators Construction to their Synthesis

Gabriele Ascari, Sara Giavante, Stefano Daddi, 2022. Applicabilità del quality-framework austriaco al Sistema Integrato dei Registri (SIR)

OECD, 2008. Handbook on Constructing Composite Indicators