Nota metodologica[1]

Metodi di imputazione di dati mancanti nelle serie storiche di indicatori utilizzati per il calcolo degli indici compositi

Dal 2017 l'Alleanza Italiana per lo Sviluppo Sostenibile (ASviS) calcola per ogni goal dell'Agenda 2030 un indice composito[2] che permette di descrivere in modo sintetico il posizionamento e l'andamento, rispetto ad ogni Goal, a livello regionale, nazionale e della Ue. Gli indici compositi sono costruiti a partire da una selezione di indicatori elementari che rispecchiano i target esplicitati all'interno di ogni goal dell'Agenda 2030. La selezione è stata effettuata dall'Area ricerca e dagli esperti dei singoli gruppi di lavoro di ASviS.

Nella raccolta e nella selezione degli indicatori elementari è possibile che le serie storiche presentino dei dati mancanti per diverse motivazioni. A partire dalle possibili casistiche, di seguito vengono esplicitati gli approcci metodologici di trattamento dei dati mancanti utilizzati. È necessario specificare che tali metodi di imputazione vengono applicati a partire dall'ipotesi che non ci siano stati eventi esterni che possano aver condizionato l'andamento delle serie storiche, causandone l'interruzione o il salto. Ne consegue che tali metodi vengono utilizzati solamente per gli anni in cui non sono noti eventi particolarmente estremi che possano aver determinato forti cambiamenti nelle serie storiche[3].

Qui di seguito le casistiche considerate e le relative scelte adottate per l'imputazione:

1. L'indagine in oggetto non viene svolta annualmente[4] oppure la serie storica è mancante di uno o più anni per tutte le unità del collettivo.
    a. I valori mancanti dell'indicatore vengono stimati attraverso l'utilizzo dell'interpolazione lineare, che, a partire da due valori noti, permette di calcolare uno o più valori compresi tra questi attraverso l'utilizzo di una funzione di una regressione lineare;
2. La prima occasione d'indagine è successiva al tempo *t* d'inizio della serie storica (primo anno preso in considerazione nel calcolo dell'indice composito).
    a. I valori dell'indicatore del primo anno disponibile ($t+n$) vengono replicati per l'anno o per gli anni precedenti mancanti nella serie storica;
3. La serie storica dell'indicatore non è ancora aggiornata all'ultimo o agli ultimi anni presi in esame nel calcolo dell'indice composito.
    a. I valori mancanti dell'indicatore base vengono calcolati attraverso l'utilizzo delle variazioni negli anni mancanti di un indicatore proxy – altamente correlato statisticamente, disponibile per gli anni mancanti dell'indicatore base e fortemente legato da un punto di vista concettuale. La variazione nell'indicatore base viene calcolata a partire dalla variazione osservata nei valori stimati dell'indicatore base

[1] La presente nota è stata realizzata dall'Area ricerca dell'ASviS, con il supporto degli esperti esterni Andrea Fasulo e Marco D. Terribili.

[2] Gli indici compositi vengono calcolati utilizzando la metodologia AMPI. Per maggiori dettagli si veda: Mazziotta, M., & Pareto, A. (2016). On a generalized non-compensatory composite index for measuring socio-economic phenomena. *Social indicators research*, 127(3), 983-1003.

[3] Si pensi, ad esempio, all'attuale condizione di pandemia causata dal COVID-19, la quale provocherà a partire dal 2020 e per gli anni successivi forti shock nelle serie storiche di molti degli indicatori utilizzati per il calcolo degli indici compositi degli obiettivi di sviluppo sostenibile.

[4] Esempio: Indicatore elementare sul *trattamento delle acque reflue*, proveniente dal *censimento delle acque per uso civile* (Istat - PSN:IST-02192) con cadenza triennale.

ottenuti attraverso un modello di regressione lineare, dove l'indicatore proxy viene utilizzato come variabile esplicativa;

   b. Qualora non fosse possibile individuare un indicatore proxy in grado di fornire informazioni sull'andamento dell'indicatore base negli anni mancanti della serie storica, i valori mancanti dell'indicatore base vengono ottenuti come predizione derivante da un modello lineare autoregressivo, laddove la bontà di adattamento della retta ai dati osservati mostri un R^2 superiore a 0,7;
   c. Qualora il modello non si adatti sufficientemente ai dati osservati, viene replicato per l'anno o per gli anni mancanti l'ultimo valore osservato.

4. Il dato non è stato diffuso per singola unità territoriale in corrispondenza di uno o più anni della serie storica poiché il campione non ha raggiunto il numero minimo di unità campionarie necessarie per la diffusione delle stime. Condizione che garantisce l'efficienza e la robustezza delle stime – data da un minor errore campionario e dalla verificabilità della condizione di normalità, necessaria alla costruzione degli intervalli confidenza – ed infine il rispetto della privacy a livello di micro-dati.
   a. Il valore mancante in corrispondenza della singola unità territoriale viene calcolato utilizzando la variazione osservata nei valori stimati dell'indicatore ottenuti attraverso un modello di regressione lineare, dove l'aggregato territoriale di livello superiore[5] a cui l'unità territoriale in oggetto appartiene viene utilizzato come variabile esplicativa del modello. Laddove non fossero disponibili i valori dell'indicatore in corrispondenza dell'aggregato territoriale corrispondente all'unità territoriale in oggetto, vengono applicate le soluzioni previste ai punti 1 e 2, in base al caso specifico.
5. Non è disponibile o non viene calcolato l'indicatore per l'aggregato, laddove invece è disponibile il valore per le singole unità territoriali che lo compongono[6].
   a. Il valore mancante dell'indicatore per l'aggregato viene ricostruito come media ponderata delle unità territoriali sulla base della specifica unità di misura (popolazione, km2, ecc.).
6. Non è disponibile o non viene calcolato l'indicatore per l'unità territoriale per l'intera serie storica, ma è disponibile per un aggregato[7], a patto che questo non sia l'aggregato di tutte le unità statistiche prese in considerazione[8].
   a. Il valore mancante delle singole unità territoriali viene ottenuto replicando per ognuna di esse il valore dell'indicatore in corrispondenza dell'aggregato territoriale.
7. Sono mancanti i valori di un indicatore per una specifica unità territoriale per l'intera serie storica, e non è disponibile un aggregato territoriale di livello superiore (punto 6).
   a. L'unità territoriale viene esclusa dal calcolo del composito.

[5] Esempio: per una Regione o Provincia italiana è utilizzato l'aggregato territoriale Nord-ovest, Nord-est, Centro, Sud o Isole.
[6] Esempio: è disponibile il valore dell'indicatore per le Province Autonome di Bolzano e Trento ma non il valore per il Trentino-Alto Adige.
[7] Esempio: è disponibile il valore dell'indicatore per il Trentino-Alto Adige ma non il valore per le Province Autonome di Bolzano e Trento.
[8] L'Italia per le Regioni italiane oppure l'Unione Europea per i paesi europei.