**DATA VALIDATION**

ACRONYMS

ESS - European Statistical System

ESS QAF - Quality Assurance Framework of the European Statistical System

EDIMBUS - Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys

GSBPM – Generic Statistical Business Process Model

GSIM – Generic Statistical Information Model

GSDEMs –Generic Statistical Data Editing Models

NA - National Accounts

NSI - National Statistical Institute

CSPA – Common Statistical Production Architecture

La validazione dei dati è un'attività volta a verificare se una combinazione di valori è un membro di un insieme di combinazioni accettabili (pagina 8 del Manuale ESS sulla Metodologia per la Validazione dei Dati).

La validazione dei dati e dei metadati è il processo che garantisce che i dati e i metadati siano puliti, corretti e utili. In questo contesto di validazione dei dati, la seguente definizione specifica sintetizza questo processo: la validazione è un'attività che verifica se una combinazione di valori è o meno un membro di un insieme di combinazioni accettabili (vedi definizioni).

La validazione dei dati e dei metadati è un compito fondamentale eseguito in tutti i settori statistici per garantire che sia i dati che i metadati siano di alta qualità.

L’insieme di “valori accettabili” può essere un insieme di valori possibili per un singolo campo. Ma, secondo questa definizione, può anche essere un insieme di combinazioni di valori validi per un record, una colonna o una collezione più ampia di dati. Sottolineiamo che l’insieme di valori accettabili non deve essere definito in modo estensivo

La validazione dei dati valuta la plausibilità dei dati: un risultato positivo non garantirà che i dati siano corretti, ma un risultato negativo garantirà che i dati siano incorretti.

**La validazione dei dati è una procedura decisionale che si conclude con l’accettazione o il rifiuto dei dati.** La procedura decisionale si basa generalmente su regole che esprimono le combinazioni accettabili di valori**. Le regole vengono applicate ai dati. Se i dati soddisfano le regole, il che significa che la combinazione espressa dalle regole non è violata, i dati sono considerati validi per l’uso finale a cui sono destinati.** Naturalmente, c’è la possibilità di utilizzare l’approccio complementare in cui le regole sono espresse in ‘forma negativa’: in questo caso, i dati vengono validati verificando che le combinazioni di valori non accettabili predefinite non si verifichino.

A volte, le regole utilizzate in una procedura di validazione sono suddivise in edizioni rigide/fatali e edizioni soft/query e i valori non accettabili sono classificati come ‘errati’ o ‘sospetti’ a seconda che non superino le edizioni rigide o le edizioni soft. Le edizioni rigide sono generalmente regole che devono necessariamente essere soddisfatte per ragioni logiche o matematiche (ad esempio, i bambini non possono essere più grandi dei loro genitori).

In caso di fallimento di una regola, i dati vengono esportati dalla procedura di validazione dei dati o contrassegnati rispettivamente e sono gestiti dal personale di correzione per correggere i valori in modo che soddisfino le regole, o i dati vengono considerati accettabili e le regole della validazione dei dati vengono aggiornate. **Il processo di validazione dei dati è una procedura iterativa basata sulla messa a punto delle regole che convergerà verso un insieme di regole considerate il set minimo di relazioni che devono necessariamente essere soddisfatte.**

When the validation fails, it may produce three types of error (the severity):

 Fatal error: the data are rejected;

 Warning: the data can be accepted, with some corrections or explanations from the data

provider;

** Information: the data are accepted**

Qui sotto riporto la correlazione col data editing, in cui si nota che la validation è una sua parte e per fare data editing puoi seguire come riferimento: Also we can take as reference for statistical data editing the GSDEMs and the GSBPM (pagina 6-7 vi sono tutti i dettagli).

**Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente**

Il **validation level** è un indice di che misura quanto un insieme di dati sia stato convalidato (livello di convalida) dalle procedure di convalida ad esso applicate.

TARGET PROCESS (vedi schema)

Nel processo aziendale target, le regole di validazione sono progettate e concordate congiuntamente a livello di gruppo di lavoro di ciascun settore statistico. Le regole di validazione che ne derivano sono documentate utilizzando standard comuni a tutti i settori, con chiare responsabilità di validazione assegnate ai diversi gruppi che partecipano al processo di produzione delle statistiche europee.

**L'utilizzo di standard comuni per le regole di validazione e i report di validazione**, combinato con le linee guida comuni per i servizi IT dell'ESS sviluppate dal progetto ESS Vision 2020 SERV, consentirà la creazione di servizi ESS condivisibili per la validazione. Gli Stati membri potranno utilizzarli su base volontaria per convalidare i dati da inviare a Eurostat.

Il processo aziendale di riferimento per la validazione nell'ESS comprende i seguenti passaggi del processo:

1. Progettazione delle strutture dati: Eurostat e gli Stati membri definiscono congiuntamente, a livello di gruppo di lavoro, la struttura e il formato dei dati da inviare dagli Stati membri a Eurostat. La struttura dei dati è documentata utilizzando standard di metadati accettati.
2. Progettazione delle regole di validazione: Una volta concordata la struttura dei dati, Eurostat e gli Stati membri definiscono congiuntamente, **a livello di gruppo di lavoro, le regole di validazione che devono essere applicate ai dati. Eurostat e gli Stati membri assegnano un livello di gravità a ciascuna regola di validazione e stabiliscono quale organizzazione è responsabile della sua applicazione**. Le regole di validazione sono documentate utilizzando uno standard ESS comune.
3. Convalida dei dati: Prima di inviare i dati a Eurostat, gli Stati membri applicano le regole di validazione di loro competenza. Quando Eurostat riceve i dati, verifica che gli Stati membri abbiano adempiuto ai loro compiti di convalida come previsto e applica le regole di validazione sotto la sua responsabilità**. Il risultato della procedura di convalida di Eurostat è un report di convalida che viene inviato agli Stati membri. Questo report di convalida segue una struttura standard. (Eurostat va da correttore e ti conferma che la procedura di convalida dei dati ha avuto successo tramite un report)**
4. Accettazione dei dati: Sulla base del report di convalida prodotto dalle procedure di convalida, Eurostat determina se i dati possono essere accettati o meno. Questa procedura di accettazione segue un processo standard e concordato. Se i dati vengono accettati, il processo aziendale di validazione si conclude. Se i dati non vengono accettati, viene inviata una richiesta di chiarimento o di rinvio dei dati da parte dello Stato membro.

LIVELLI DI VALIDAZIONE: esistono 6 livelli di validazione da 0 a 5 ora li elenco e descrivo in cosa consistono:

**Validation level 0**: consistency with the expected IT structural requirements

At this level, it is checked the consistency of the data with their expected IT requirements, for

instance

- if the file has been sent/prepared by the authorised authority (data sender);

- if the column separator, the end of record symbol are correctly used;

- if the file has the expected number of columns (agreed format of the file);

- if the column have the expected format of the data (i.e., alphanumeric, numeric, etc.)

For these quality checks only the structure of the file or the format of the variables are necessary

as input. (Dunque riguarda i requisiti strutturali dei dati e il formato del file).

**Validation level 1:** consistency within the data set

It is checked the consistency within the elements of the data set. For these quality checks, it is

needed only the (statistical) information included in the file itself.

For instance:

- check whether the number included in column 4 is not negative (as expected);

- check whether the year in the second column is 2011, as in the file name;

- check whether the content of the third column is one of the codes of the dictionary "Sex";

- check whether the content of the first column is consistent with the data sender (let's

assume that there is a dictionary including the list of the data senders associated to the

specific data set): data for Luxembourg should not be sent by another country.

- based on information available before data collection (for example from previous survey

or other sources) one could establish a "plausibility range" for a certain variable (for

instance number of components of a household).

- check consistency at (micro-level) of two (or more) variables: a certain combination of

codes is illogical, a variable has to be reported only for a certain combination of codes.

- check consistency at macro-level of two (or more) variables: Total inhabitants = male

inhabitants + female inhabitants, or Female inhabitants = (total inhabitants / 2)

Livello 1 quindi se la colonna 4 inerente all’età contiene tutti numeri positivi ad esempio.

**Validation level 2:** coerenza con altri set di dati dello stesso dominio e della stessa fonte di dati Il livello di convalida 2 si occupa della verifica della coerenza basata sul confronto del contenuto del file con il contenuto di "altri file" che si riferiscono allo stesso sistema statistico (o dominio) e alla stessa fonte di dati. Ad esempio: Caso a) gli "altri file" possono essere altre versioni esattamente dello stesso file. In questo caso, i controlli di qualità sono destinati a rilevare "revisioni" rispetto ai dati precedentemente inviati. La rilevazione e l'analisi delle revisioni possono essere utili, ad esempio, per verificare se le revisioni sono coerenti con gli outlier rilevati nei precedenti controlli di qualità (correzioni) o per avere una stima dell'impatto delle revisioni nei risultati "da pubblicare", a beneficio degli utenti. Caso b) gli "altri file" possono essere versioni dello stesso set di dati riferite ad altri periodi temporali. Questi controlli sono comunemente chiamati "controlli delle serie temporali" e sono destinati a verificare la plausibilità delle serie temporali. Caso c) gli "altri file" possono fare riferimento ad altri set di dati dello stesso fornitore di dati (ad esempio, Paesi nell'ESS), facendo riferimento agli stessi o ad altri periodi temporali correlati. A volte un gruppo di set di dati (stesso paese, stesso periodo di riferimento) viene inviato contemporaneamente. Esempio: potrebbero essere inviati contemporaneamente tre file, dello stesso paese e relativi allo stesso periodo di tempo: un file include dati per "femmine", uno per "maschi" e uno per "totale". La coerenza tra i risultati dei tre file può essere verificata. Altro esempio: i risultati dai set di dati annuali possono essere confrontati con i risultati dei corrispondenti dati trimestrali.

**Livello di convalida 3**: coerenza all'interno dello stesso dominio tra diverse fonti di dati Il livello di convalida 3 si occupa della verifica della coerenza basata sul confronto del contenuto del file con il contenuto di "altri file" che fanno riferimento a un diverso fornitore di dati sullo stesso sistema statistico o dominio armonizzato (condividendo standard comuni riguardo a ambito, definizioni, unità e classificazioni nelle diverse indagini e fonti). Ad esempio: Caso d) gli "altri file" possono fare riferimento allo stesso set di dati, ma provenienti da un altro fornitore di dati (ad esempio, paesi dell'ESS, che condividono metodologie armonizzate). In questa classe sono inclusi anche i controlli specchio. Le "statistiche specchio coinvolgono questioni di coerenza, comparabilità geografica nonché precisione" (Eurostat, 2014). Spesso, tali statistiche sono importanti per l'analisi dei dati a livello di Eurostat. I controlli specchio verificano la coerenza tra dichiarazioni provenienti da diverse fonti che si riferiscono allo stesso fenomeno, ad esempio, le esportazioni dichiarate dal paese A al paese B dovrebbero essere le stesse delle importazioni dichiarate dal paese B provenienti dal paese A.

**Livello di convalida 4**: coerenza tra domini separati nella stessa fonte di dati Il livello di convalida 4 potrebbe essere definito come controlli di plausibilità o coerenza tra domini separati disponibili nella stessa istituzione. La disponibilità implica un certo livello di "controllo" sulle metodologie da parte dell'istituzione interessata. Questi controlli potrebbero basarsi sulla plausibilità dei risultati che descrivono lo "stesso" fenomeno da diversi domini statistici. Esempi: disoccupazione da registri e dall'Indagine sulle forze di lavoro, o abitabilità di un'abitazione (dall'indagine sui proprietari di case e abitazioni rispetto al registro della popolazione). Controlli potrebbero essere effettuati anche tra risultati provenienti da fonti di microdati e macrodati correlate. Altri controlli di plausibilità potrebbero basarsi su correlazioni note tra diversi fenomeni: ad esempio, commercio esterno e attività di trasporto internazionale in porto.

**Livello 5:** Validation level 5 could be defined as plausibility or consistency checks between the data available

in the data provider (Institution) and the data / information available outside the data provider

(Institution). This implies no "control" over the methodology on the basis of which the external

data are collected, and sometimes a limited knowledge of it.

Esempi di **structural validation rules**: Examples of rules to ensure technical integrity of a data file (Structural Validation)

Formal validity of…

- data type Telephone number: Numerical data.

- field length Date: If Date is given as text it should be 8 characters long

- characters Date: If Date is given as text it should contain only numbers.

- numerical range Month: Month of arrival in the country must be in {1,...,12}

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamenteTABELLA QUI sotto esempi di Structural vaidation (es: no duplicati per l’ID number di uno studente)

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

**8 The data validation process life cycle**

In order to improve the performance of a statistical production process by managing and

optimising the data validation process, it is useful to describe the data validation process life cycle.

Questo processo è costituito da 4 fasi: cycle starts by **designing** the data validation

process for the statistical domain or inter statistical domain, with an overall study of the data sets,

variables and its relations to find a list of suitable and effective validation rules. In the

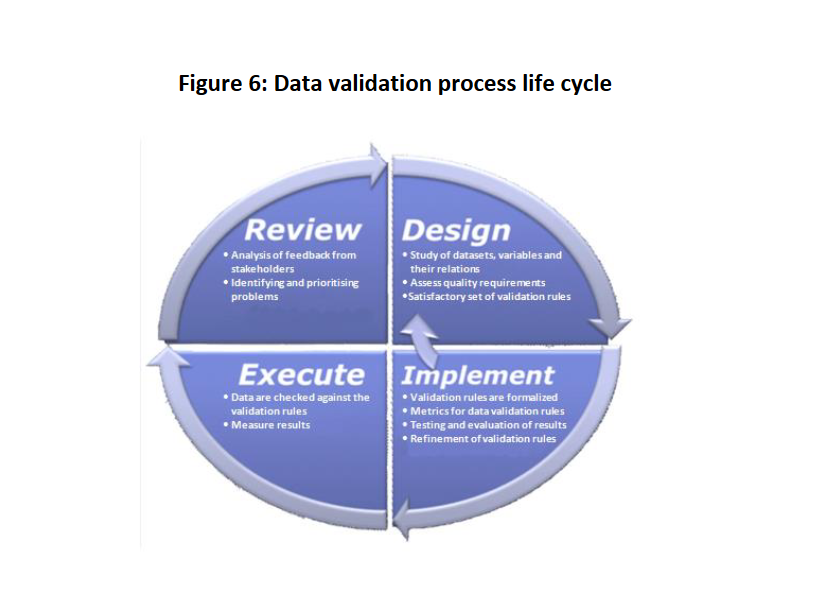
**implementation** phase, these validation rules are described in common syntax, formalised, tested

and refined, discussed and evaluated by stakeholders. During the **execution** phase, data are

checked against the rules; validation results are measured and quantified. These outputs will be

**reviewed** to improve the list of validation rules.

Quindi si studia il dataset e le relazioni tra le variabili, si implementano le validation rules dopo averle discusse, si applicano tali rules al dataset misurando e quantificando i risultati ed infine si fa una review di questi migliorando le regole ove possibile applicandole di nuovo eventualmente. (quindi è iterativo)



FASE DI DESIGN

. La progettazione di un insieme di regole di convalida è un processo dinamico. Le regole di convalida dovrebbero essere progettate in collaborazione con specialisti del settore e dovrebbero basarsi sull'analisi di indagini precedenti. La coerenza e la non ridondanza delle regole dovrebbero essere verificate. Le regole di convalida dovrebbero essere progettate con cautela al fine di evitare eccessive modifiche. Regole di convalida efficaci possono essere ottenute combinando in modo diverso approcci e "best practices".

In questa fase, il processo di convalida dovrebbe essere pianificato e documentato per un ulteriore monitoraggio del progresso. Dovrebbe essere considerata la gestione complessiva del processo e le interfacce con gli altri sottoprocessi. Per ogni fase, è necessario pianificare le risorse e il tempo necessari per implementare, testare, eseguire, revisionare e documentare.

Questa è la fase in cui progettisti di indagini, progettisti di questionari, specialisti di convalida e modifica, e esperti del settore devono cooperare.

Descrizioni delle attività:

* Valutare i requisiti di qualità per i set di dati.
* Studio complessivo dei set di dati, delle variabili e delle loro relazioni.
* Determinare un insieme soddisfacente di regole di convalida per i dati, al fine di rendere il processo di produzione dei dati più efficiente, riducendo il tempo e le risorse umane, ma considerando i requisiti di qualità.
* Valutare responsabilità e ruoli. Documentare chi fa cosa; chi è responsabile di diverse azioni; chi accetta e adotta le regole di convalida, ecc.
* Integrare il processo di convalida dei dati nell'intero processo di produzione statistica. Progettare le connessioni con le altre fasi dei processi di produzione statistica.
* Migliorare la convalida in base ai risultati della fase di revisione. Un documento con la forma di linee guida con alcuni fondamenti teorici, esempi e migliori pratiche potrebbe supportare il compito del responsabile di dominio nella progettazione dell'intero processo di convalida.

FASE DI IMPLEMENTAZIONE

È importante capire che una volta implementato il processo di convalida nel processo di indagine effettivo, dovrebbero essere apportate solo leggere modifiche al monitoraggio e al taratura al fine di evitare cambiamenti strutturali.

Descrizioni delle attività: (validation rules)

* Formalizzazione e descrizione delle regole di convalida in una sintassi comune.
* Determinazione delle metriche per le regole di convalida dei dati, la valutazione del processo di convalida e le regole di convalida. Le regole di convalida dovrebbero essere valutate per la qualità (chiare, inequivocabili e coerenti, risparmiando risorse temporali).
* Testing. Applicare le regole di convalida ai dati di prova (dati reali, dati artificiali) e produrre indicatori.
* I risultati del test (indicatori, regole di convalida, metriche, aspetti qualitativi, ecc.) sono valutati dagli interessati (Eurostat, Stati membri, responsabili di dominio, ecc.). Documenti di report sui risultati del test e sulla valutazione dovrebbero essere preparati e salvati per la fase di revisione.
* Perfezionamento delle regole di convalida in base ai risultati del test e alle consultazioni con gli interessati.
* Documentazione. Le regole di convalida dei dati dovrebbero essere ben documentate: i documenti dipendono dall'obiettivo e dall'utente finale, che siano produttori, utenti dei risultati, gestori di indagini o metodologi.

FASE DI ESECUZIONE

La fase di esecuzione consiste nell'identificare i valori che non sono accettabili rispetto alle validation rules che consistono in relazioni logiche, matematiche o statistiche. Questo processo di solito comprende un insieme di metodi di convalida integrati che trattano diversi tipi di errori. Ciò consente di valutare la qualità dei dati e aiuta a identificare le fonti di errore per future migliorie del processo di produzione statistica. Lo scopo di questa fase è raccogliere le statistiche sui risultati della convalida per valutare la qualità dei set di dati e la qualità delle regole di convalida

Quindi consiste nella verifica dei risultati e identificazione dei dati che non vanno bene nel dataset.

Il risultato della fase di esecuzione è un indicatore che segnala dati accettabili e non accettabili e, generalmente, un punteggio che misura il grado di gravità del fallimento. Una comunicazione standardizzata di messaggi di errore/avviso può aumentare l'efficienza complessiva della produzione statistica e influisce direttamente sul tempo necessario per comprendere e individuare la fonte dell'errore.

Ulteriori informazioni sulle metriche di convalida possono essere trovate nella seconda parte di questo manuale: Metriche per una procedura di convalida dei dati!!! Capitolo 9 metriche per la convalida dei dati (se i va ben o meno).

**Activity descriptions:**

 Data are checked against the validation rules. Validate data against predefined validation

rules.

 Summarising results. It depends on the user of the results (staff, management or

methodologist)

Quindi check dei dati quali vanno bene o no e si riassumono i risultati, magari con le metriche del capitolo 9-10.

FASE DI REVIEW

Questa fase mira al miglioramento continuo dell'efficacia del processo di convalida e della qualità dei dati. Durante la fase di revisione vengono stabiliti i requisiti per nuovi elementi di progettazione. Questa fase comprende l'identificazione dei problemi utilizzando l'analisi dei risultati della fase di esecuzione. I problemi identificati vengono prioritizzati e affrontati nella fase di progettazione.

Esempi di revisioni includono:

* Miglioramento delle regole di convalida a causa di:
  + Sostituzione di quelle che rilevano pochi errori con altre più potenti.
  + Sostituzione di quelle che "traggono in inganno": rilevano errori che non sono veri errori.
  + Aumento dell'efficienza delle regole di convalida.
  + Miglioramenti nelle regole di convalida: rilevare più possibili errori.
  + Modifiche nel file dati o nelle normative.
* Cambiamenti nel processo di convalida originati da:
  + Cambiamenti negli strumenti di convalida.
  + Cambiamenti nei formati dei file.
  + Miglioramento dell'efficienza.
* Cambiamenti nel flusso di lavoro di convalida a causa di:
  + Migliore assegnazione delle responsabilità nelle attività di convalida.
  + Guadagni di efficienza nella catena.

Descrizioni delle attività:

* Analisi dei feedback degli interessati. Feedback raccolto nelle fasi precedenti.
* Analisi dei risultati della fase di esecuzione. Problemi potenziali identificati, errori, discrepanze, problemi sistematici rilevati vengono analizzati per decidere se le regole di convalida dovrebbero essere riviste.
* Identificazione e prioritizzazione dei problemi.

**Praticamente migliori le regole di validazione.**

CAPITOLO 10 Metriche per la valutazione della data validation

Per facilitare tali valutazioni, è utile poter quantificare le prestazioni attuali della procedura in qualche modo. Esempi includono il conteggio delle violazioni per regola o per record, il numero di regole utilizzate in una procedura (poiché tendono a proliferare nella pratica), il numero di regole ridondanti e così via. Qualsiasi tale quantificazione è una metrica (indicatore di qualità) della procedura di convalida dei dati**. Le metriche dovrebbero misurare l'efficacia e l'efficienza di una procedura di convalida dei dati.** **Gli indicatori possono riferirsi sia a una singola regola che all'intera procedura di convalida dei dati**. Le metriche che saranno discusse nelle prossime sottosezioni possono essere distinte in:

* Indicatori che tengono conto solo delle regole di convalida (proprietà delle regole di convalida)
* Indicatori che tengono conto solo dei dati osservati
* Indicatori che tengono conto sia dei dati osservati che dei dati di riferimento (ad esempio, dati imputati, dati simulati).

The first two are generally used to fine tune the data validation procedure, for instance in the

design phase and by using pilot survey.

The metrics of the third type are used to obtain a more precise measure of the effectiveness of a data validation procedure.