MODULO PROBABILISTICO PER L'INDIVIDUAZIONE DEI RILANCI DA SCANNER DATA

Sommario

Introduzione

Il modulo è pensato per elaborare due file CSV, uno di prodotti usciti dal mercato ed uno di nuovi prodotti entrati in anagrafica, per individuare i *"rilanci"*, ovvero quei prodotti che vengono sostituiti con nuove edizioni simili e che quindi possono concorrere come coppia al calcolo dell'andamento degli indici dei prezzi.

Esprimiamo il rilancio con una coppia di prodotti, uno uscito dal mercato (presente nel primo file) ed uno entrato nel mercato (presente nel secondo file).

Il modulo individua i rilanci con un metodo probabilistico che si basa sulla somiglianza delle variabili descrittive presenti nei due file. Il modulo utilizza il software di record linkage MEARLIN realizzato in R.

Requisiti per l'esecuzione del modulo

Per eseguire il modulo dei rilanci è necessario che sulla macchina sia installato **R**.

Inoltre la variabile di sistema "**PATH**" deve contenere il riferimento alla directory degli eseguibili di R (ad esempio qualcosa tipo "C:\Program Files\R\R-3.6.1\bin"). In alternativa deve essere modificato lo script "**Rilanci.bat**" inserendo qui il percorso locale di R.

È necessario installare i pacchetti utilizzati dal software MAERLIN ovvero:

- **proxy**
- **tm**
- **RecordLinkage**

Ciò può essere fatto digitando il comando seguente in una sessione R interattiva:

```
install.packages(c("tm", "proxy", "RecordLinkage"))
```

È necessario che i file CSV dei prodotti contengano le seguenti variabili: '**PRDKEY', 'MERCATO', 'MARCA', 'DESCRIZIONE', 'FORMATO', 'NUM_PEZZI'**.

Descrizione del pacchetto

Il pacchetto di rilascio del modulo contiene quattro cartelle:

1) **Application**: Questa cartella contiene il file di parametri (vedi sotto) e gli script di lancio del modulo (quello da eseguire è "Rilanci.bat").
2) **Input**: in questa cartella devono essere spostati i file da elaborare (quello dei prodotti usciti e quello dei prodotti entrati); si consiglia di creare delle sotto cartelle una per ogni mese di elaborazione in modo da poter mantenere uno storico ordinato delle elaborazioni delle mensilità precedenti.

3) **Output**: in questa cartella saranno generati i file di log e i file con i risultati dell'elaborazione; anche in questo caso si consiglia di creare delle sotto cartelle una per ogni mese di elaborazione (le sotto cartelle saranno generate direttamente dal modulo se indicato nel file di parametri, vedi sotto).
4) **Sources**: questa cartella contiene tutti i sorgenti R del software MEARLIN e non deve essere modificata.

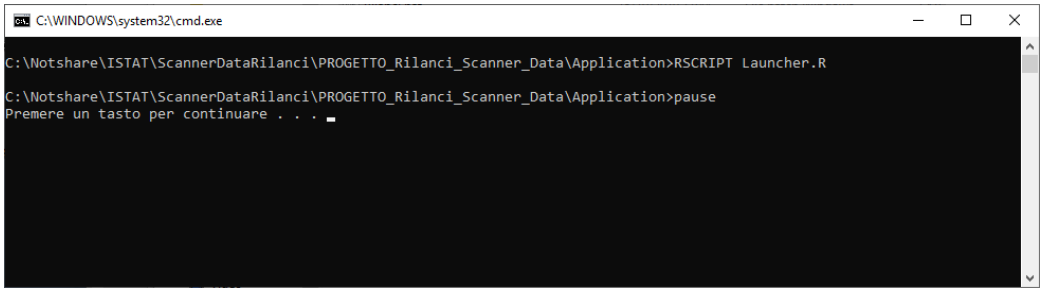
Come si lancia il modulo

Eseguire il modulo di individuazione dei rilanci è molto semplice:

1) spostare i file da elaborare sotto la cartella "**Input**" (si consiglia di creare sottocartelle per anno e mese di riferimento dei dati);
2) accedere alla cartella "**Application**";
3) compilare il file "**file.param**" (descritto sotto più dettagliatamente) indicando i riferimenti dei file da elaborare e delle cartelle di lavoro;
4) eseguile lo script "**Rilanci.bat**";
5) spostare i file **ID_Links_ACCETTATI.csv** e **Links_ACCETTATI_ELAB.csv** della cartella di output alla directory prevista per l'acquisizione da parte della procedura di calcolo dell'indice dei prezzi. Il primo file contiene i rilanci individuati dal modulo, indicati con una coppia di PRDKEY, nel secondo file ci sono alcune informazione relative ai prodotti oggetto di rilancio utili nelle fasi successive dell'elaborazione.

Se i file non vengono creati si può consultare il file di log per una diagnosi del problema. Se non ci sono errori il modulo genererà una serie di output (descritti sotto più dettagliatamente) tra cui appunto i due file da spostare.

Il lancio dello script aprirà una finestra di comando che, al termine dell'elaborazione, apparirà come nella figura.



L'elaborazione completa dura **pochi minuti**.

Compilazione del file di parametri file.param

Di seguito sono elencati tutti i parametri da editare in "**file.param**":

- **InputDir_ANNO_MESE**: nome della cartella dove si trovano i file da elaborare; il percorso è relativo alla cartella dove si trova il file dei parametri; il valore del parametro deve essere indicato tra doppi apici e il

carattere backslash deve sempre essere raddoppiato; la cartella deve essere contenuta nella directory "**Input**"
(Esempio: InputDir_ANNO_MESE <- "..\\Input\\202001").

- **FileA**: nome del file di dati contenente la lista dei prodotti usciti dal mercato
(Esempio: FileA <- "Uscenti.csv").
- **FileB**: nome del file di dati contenente la lista dei prodotti nuovi entrati nel mercato
(Esempio: FileB <- "Anagrafica.csv").
- **OutputDir_ANNO_MESE**: nome della cartella dove verranno generati i file dei risultati ed il log; il percorso è relativo alla cartella dove si trova il file dei parametri; il valore del parametro deve essere indicato tra doppi apici e il carattere backslash deve sempre essere raddoppiato; la cartella deve essere contenuta nella directory "**Output**"
(Esempio: OutputDir_ANNO_MESE <- "..\\Output\\202001")
- **Dir_Packages**: nome della cartella contenente i packages R richiesti; se questo parametro è omesso il modulo cercherà i packages nella directory di default dell'utente, il percorso può essere relativo o assoluto; il valore del parametro deve essere indicato tra doppi apici e il carattere backslash deve sempre essere raddoppiato.
(Esempio: Dir_Packages <- "..\\..\\3.6")

Regole deterministiche

I colleghi di produzione hanno fornito delle regole deterministiche di "rifiuto" di rilanci potenziali individuati dal modello probabilistico. La lista delle regole è:

1) I due prodotti del rilancio non possono avere valori diversi nella unità di misura della variabile FORMATO.
2) I due prodotti del rilancio non possono avere una differenza relativa nella quantità della variabile FORMATO superiore al 25%.
3) I due prodotti del rilancio non possono avere valore mancante nella quantità della variabile FORMATO.
4) I due prodotti del rilancio non possono avere valori diversi nella variabile MARCA
5) I due prodotti del rilancio non possono avere valori diversi nella variabile NUM_PEZZI

Risultati dell'esecuzione del modulo

Il modulo crea una serie di file di testo nella cartella indicata dal parametro "**OutputDir_ANNO_MESE**" specificato in **file.param**.

Alcuni **file** sono **per la consultazione degli utenti**. La struttura di questi file è la seguente: ogni coppia di record che rappresenta un rilancio è riportata su due righe consecutive: la prima riga si riferisce al prodotto uscito, la seconda al prodotto entrante abbinato. Il tracciato record di questi file riporta i campi descrittivi dei prodotti ed il campo '**fm.d**' che fornisce la probabilità a posteriori stimata dal modello che la coppia sia un rilancio.

Altri **file** sono **in formato sintetico** e pensati per essere successivamente elaborati. La struttura di questi file è tabellare ed ogni riga rappresenta un rilancio. Le colonne della tabella si chiamano 'PRDKEY.A' e 'PRDKEY.B' e forniscono rispettivamente gli ID dei prodotti uscenti ed entranti abbinati come rilancio. **NOTA: Su richiesta dei colleghi DCIT i nomi delle colonne sono stati eliminati dai file sintetici (17/01/2020). La semantica delle colonne rimane quella qui documentata.**

Un formato particolare invece è previsto per l'ultimo file: **Links_ACCETTATI_ELAB.csv.** Ogni coppia di record che rappresenta un rilancio è riportata su due righe consecutive, i valori nei campi non sono mai delimitati da apici o doppi apici e i nomi delle colonne sono stati eliminati. Il tracciato record rimane quello di seguito documentato:

FILE, PRDKEY, YEAR, MONTH, INDICODECR, COICOP, UNITA, QUANTITA

Segue l'elenco dei file ed una breve descrizione:

- **Links.csv**: File per la *consultazione* con tutti i rilanci individuati dal modello probabilistico prima dell'imposizione delle regole deterministiche.
- **Links_SCARTATI.csv**: File per la *consultazione* con i rilanci individuati dal modello scartati perché non verificano le regole deterministiche. La colonna "REGOLA" riporta l'identificativo numerico della regola responsabile dell'esclusione secondo la numerazione riportata nel paragrafo "Regole deterministiche".
- **Links_ACCETTATI.csv**: File per la *consultazione* con i rilanci individuati dal modello che verificano le regole deterministiche.
- **ID_Links.csv**: File *sintetico* con tutti i rilanci individuati dal modello probabilistico prima dell'imposizione delle regole deterministiche.
- **ID_Links_SCARTATI.csv**: File *sintetico* con i rilanci individuati dal modello scartati perché non verificano le regole deterministiche.
- **ID_Links_ACCETTATI.csv**: File *sintetico* con i rilanci individuati dal modello che verificano le regole deterministiche; **questo è uno dei file che deve essere spostato ed acquisito dalle successive fasi del calcolo**.
- **Links_ACCETTATI_ELAB.csv:** File con dati di dettaglio relativi ai prodotti oggetto di rilancio; **questo è uno dei file che deve essere spostato ed acquisito dalle successive fasi del calcolo**.
- **LogRilanci.txt**: *File di log* dell'elaborazione effettuata dal modulo.
- **LogRilanciMsg.txt**: *File di log* dell'elaborazione effettuata dal modulo completo di messaggistiche di errore da consultare in caso l'elaborazione non si sia conclusa correttamente.
- **MAERLIN.RData**: Salvataggio dell'area di memoria di R relativa all'elaborazione, da consultare con R in caso di necessità per analisi dettagliate.

Riferimenti bibliografici sulla metodologia utilizzata

D. Zardetto, M. Scannapieco (2010) **"A Novel Suite of Methods for Mixture Based Record Linkage"**, in: *"Rivista di Statistica Ufficiale"*, n. 2-3/2010, Istat, pagg. 31-58, ISSN 1828-1982.
URL: https://www.istat.it/it/files/2011/09/rivista_statistica_ufficiale_2_3_2010.pdf

D. Zardetto, L. Valentino, M. Scannapieco (2011) **"MAERLIN: New Record Linkage Methods At Work"**, in: *"Proceedings of the 6th International Conference on New Techniques and Technologies for Statistics (NTTS 2011)"*, Bruxelles, Belgio, 22-24 Febbraio 2011.
URL: https://ec.europa.eu/eurostat/cros/system/files/PS4%20Poster%202.pdf

D. Zardetto, M. Scannapieco, L. Valentino, T. Catarci (2011) **"On Probabilistic Record Linkage: New Methods Compared to the Fellegi-Sunter Approach"**, in: *"Proceedings of the 19th Italian Symposium on Advanced Database Systems (SEBD 2011)"*, pagg. 21-32, Maratea, Italia, 26-29 Giugno 2011.
URL: http://www.sebd.org/2011/documents/SEBD2011-Proceedings.pdf