Problemi di Metodologia Statistica

Ordine degli interventi:

1. Le basi metodologiche del record linkage

   Cristina Mazzali
   Università degli Studi di Milano

2. La scelta degli strumenti di record linkage

   Sara Poidomani
   Azienda Ospedaliera "Ospedale di Circolo di Melegnano"

3. La funzione "Soundex"

   Paolo Borsa
   Metodologie e Tecniche di Comunicazione Linguistica – Politecnico di Milano

Problemi di Metodologia Statistica

3. La funzione ‘Soundex’

Paolo Borsa

Metodologie e Tecniche di Comunicazione Linguistica –

Politecnico di Milano

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- **definizione**
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- sviluppi possibili

Soundex è un **algoritmo fonetico** per indicizzare nomi in base al loro suono

[brevetto Robert C. Russel, 1918 e 1922]

Il suo scopo principale è quello di consentire la **transcodifica in una medesima stringa** di nomi con pronuncia uguale ma diversa rappresentazione grafica

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- **utilizzo**
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- sviluppi possibili

Un algoritmo di ricerca Soundex prende una parola – generalmente un nome di persona – come input e produce una stringa di caratteri alfanumerici che identifica un gruppo di parole la cui realizzazione fonetica è (più o meno) simile.

È utile per condurre ricerche in ampi database, allorché si posseggono dati incompleti, disomogenei, corrotti.

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- **utilizzo**
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- sviluppi possibili

La Soundex è stata sviluppata per la **lingua inglese**, nella quale il rapporto di corrispondenza tra suoni (*fonemi*: unità foniche con valore distintivo) e segni (*grafemi*: unità di scrittura) è piuttosto libero, e per il contesto statunitense, in cui i nomi propri hanno origine etnica diversa.

Consente di collegare tra loro, da database diversi, nomi con ortografia multipla o mobile, o con **errori di ortografia**.

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- **descrizione**
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- sviluppi possibili

Il codice Soundex è **una lettera seguita da tre numeri**:

la lettera è la prima lettera del nome,

i numeri transcodificano le rimanenti consonanti.

L'alfabeto di riferimento è quello della lingua inglese (26 grafemi).

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- **descrizione**
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- sviluppi possibili

L'algoritmo:

1) Eliminare dal nome i segni interpuntivi, i segni diacritici e gli spazi
2) Rimuovere le seguenti lettere, salvo il caso in cui si tratti della lettera inziale: A, E, I, O, U, Y, W, H
3) Mantenere la prima lettera del nome e trasformare le altre in cifre, secondo la seguente tabella:

1	←	B, P, F, V	labiali, labiodentali
2	←	C, K, G, J, Q, S, Z, X	velari, sibilanti
3	←	D, T	dentali
4	←	L	laterale
5	←	M, N	nasali
6	←	R	postalveolare

4) Scempiare le coppie di cifre uguali che risultano ora accostate nella stringa
5) Completare eventualmente la stringa con 0

Problemi di Metodologia Statistica

1 ← B, P, F, V

2 ← C, K, G, J, Q, S, Z, X

3 ← D, T

4 ← L

5 ← M, N

6 ← R

Esempi:

Mr. **Allan Holdsworth**

ALLAN	A445	→	A450
HOLDSWORTH	H43263	→	H432

Mr. **Herman Ashcroft**

HERMAN	H655	→	H650
ASHCROFT	A22613	→	A261

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- **sviluppi e limiti**
- altre funzioni
- sperimentazione
- inglese e italiano
- sviluppi possibili

Una variante perfezionata dell'algoritmo è la cosiddetta **'Census Soundex'**, utilizzata nei censimenti americani a partire già dal 1880 (transcodificazione a mano).

Essa prevede che i sei grafemi identificativi delle vocali (A, E, I, O, U, Y) **non cadano** immediatamente (al passaggio 2), ma siano commutate in 0, rimanendo come **separatori**, e vengano eliminate solo dopo il passaggio 4 (scempiamento delle coppie di cifre uguali), e prima dell'eventuale completamento della stringa con 0.

1 ← B, P, F, V

2 ← C, K, G, J, Q, S, Z, X

3 ← D, T

4 ← L

5 ← M, N

6 ← R

Metodo 'classico':

Mr. **Herman Ashcroft**

HERMAN	H655	→	H650
ASHCROFT	A22613	→	A261

Metodo 'Census':

Mr. **Herman Ashcroft**

HERMAN	H06505	→	**H655**
ASHCROFT	A226013	→	A261
	A2026013	→	**A221**

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- **sviluppi e limiti**
- altre funzioni
- sperimentazione
- inglese e italiano
- sviluppi possibili

Anche se generalmente efficace, il metodo sfruttato dall'algoritmo Soundex è lungi dal costituire una soluzione perfetta. Alcuni problemi:

- Conservazione della lettera iniziale; se registrata in modo errato inficia in partenza la catalogazione del nome
- Nessi consonantici:

  -DG-
  -GH-, -GHT-, -GN-
  -KN-
  -MB-, -MP- (seguito da S, Z, T)
  -NG-, -NGT-, -NGHT-
  PF-, -PH-, PS-
  -TCH-

- Caduta di W quando ha suono bilabiale
  (es. Greensworo)

Problemi di Metodologia Statistica

1 ← B, P, F, V

2 ← C, K, G, J, Q, S, Z, X

3 ← D, T

4 ← L

5 ← M, N

6 ← R

Anche se generalmente efficace, il metodo sfruttato dall'algoritmo Soundex è lungi dal costituire una soluzione perfetta. Alcuni problemi:

- Conservazione della lettera iniziale; se registrata in modo errato inficia in partenza la catalogazione del nome
- Nessi consonantici:

Mr. **KNIGHT**	K523
KNITE	K530
NIGHT	N230
NITE	N300

- Caduta di W quando ha suono bilabiale (es. Greensworo)

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- **altre funzioni**
- sperimentazione
- inglese e italiano
- sviluppi possibili

Per rendere più preciso ed efficace il processo di indicizzazione dei nomi, sono stati messi a punto metodi più affinati e algoritmi più complessi.

A parte la 'Reverse Soundex' (variante dell'algoritmo-base, che premette alla stringa l'ultima lettera del nome), si tratta però di metodi validi essenzialmente per **contesti e/o utenti anglofoni**:

- NYSIIS algorithm
- Celko Improved Soundex (Joe Celko)
- Daitch-Mokotoff Soundex ('Jewish Soundex' o 'Eastern European Soundex')
- Metaphone e Double-Metaphone (Lawrence Philips)

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- **sperimentazione**
- inglese e italiano
- sviluppi possibili

Tra le molte soluzioni a disposizione, l'algoritmo Soundex risulta quello maggiormente adattabile a contesti non-anglofoni (in cui si utilizzi l'alfabeto latino), proprio in virtù della sua **scarsa sofisticazione**.

Nella Soundex la transcodifica delle lettere in cifre avviene, infatti, in base a una suddivisione piuttosto grezza dei grafemi e dei fonemi corrispondenti:

POCHE REGOLE ⟶ **GRANDE DUTTILITÀ**

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- **sperimentazione**
- inglese e italiano
- sviluppi possibili

Nel caso della nostra sperimentazione, inoltre, la scelta della Soundex è stata determinata anche dalla peculiare natura dei record in una delle due banche dati (SDO), sotto forma di **codice fiscale** e, dunque, con cognomi e nomi già ridotti alla loro sostanza consonantica
(salvo casi sporadici, ad es. Ugo Fedi → FDEGUO).

Per procedere al legame è stato necessario ricondurre anche le anagrafiche della seconda banca dati (ReNCaM) alla 'forma-CodiceFiscale', applicando poi la Soundex secondo la formula classica (*non-separating vowels*).

L'adozione di un algoritmo poco sofisticato si è rivelata una soluzione funzionale; i suoi limiti sono stati compensati nella fase di blocking dall'uso di ulteriori variabili (ad es. il sesso).

Problemi di Metodologia Statistica

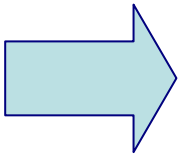
3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- **sperimentazione**
- inglese e italiano
- sviluppi possibili

Benché in casi particolari sia possibile fare ricorso senza troppi problemi agli strumenti già disponibili, approntati per la lingua inglese, tuttavia appare oggi necessario mettere a punto un metodo specifico **per la lingua italiana**, di cui vi è grande richiesta ma ancora inesistente.

Accade infatti spesso di dover lavorare su:

- banche dati molto grandi
- banche dati con record
  - ✓ poco/non controllati
  - ✓ scorretti (reg. orale o a mano)
  - ✓ disomogenei



criticità della funzione 'Soundex'

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- **sperimentazione**
- inglese e italiano
- sviluppi possibili

Per una 'Soundex_Ita'

Studio-pilota che porti prima all'elaborazione di una **Soundex per l'italiano** (e poi, eventualmente, di un metodo più sofisticato, indipendente dall'algoritmo considerato)

↓

Modifica dell'algoritmo originale (in particolare attraverso una riconsiderazione dei gruppi di consonanti), tenendo conto delle specifiche caratteristiche fonologiche dell'italiano rispetto all'inglese

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- **inglese e italiano**
- sviluppi possibili

Rispetto all'inglese, in italiano vi è una **minore libertà nel rapporto tra fonemi e grafemi**, ossia tra pronuncia e scrittura (anche se non vi è corrispondenza biunivoca)

- Minori problemi di trascrizione
- Facilità nel ricondurre a scritture corrette anche pronunce regionali marcate

Nella trascrizione i problemi non si manifestano per incertezze dovute al sistema grafematico, ma solo in presenza di trascrittori incolti o distratti, che incorrono in veri e propri **errori ortografici**

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- **inglese e italiano**
- sviluppi possibili

In italiano:

- il grafema Z rappresenta l'affricata alveolare sorda e sonora (Zeno, Graziano), mentre in inglese identifica la continua costrittiva alveolare sonora (Zachariah)
- il grafema G (+ E, I) rappresenta l'affricata prepalatale sonora /dʒ/ (Gerardo, Giorgio; ingl. Gerard, George), ma in inglese identifica spesso anche una continua costrittiva alveolare sorda (Gilbert)
- il grafema C (+ E, I) rappresenta l'affricata prepalatale sorda /tʃ/ (Cesare, Ciro), mentre in inglese identifica una continua costrittiva alveolare sorda (Cesar, Cirus)

→ riconsiderazione del gruppo di consonanti transcodificate con la cifra 2

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- **inglese e italiano**
- sviluppi possibili

Inoltre:

- le consonanti doppie hanno valore fonematico, cioè distintivo (papa : pappa, tufo : tuffo, m'ama: mamma)
- il grafema SC (+ E, I) identifica la continua costrittiva prepalatale sorda /ʃ/ (scena, sciame)
- i grafemi CH e GH (+ E, I) identificano l'occlusiva velare sorda e sonora (chilo, ghiro)
- il grafema GN identifica l'occlusiva palatale sonora nasale /ɲ/ (gnocchi, legno)
- il grafema GL (+ I) / GLI (+ A, E, O, U) identifica la continua laterale palatale /ʎ/ (gli, taglio)

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- **sviluppi possibili**

Proposte:

- **Separare** le **velari** C, K, G, J, Q dalle **alveolari** (o sibilanti) costrittive e affricate S e Z, accludendo il grafema X, che rappresenta due fonemi pronunciati in rapida successione, alle alveolari (Alexia / Alessia)

- **Accorpare** in un unico gruppo le **liquide** L e R, in considerazione del fatto che quest'ultima è in italiano una vibrante (nei singoli parlanti può anche essere blesa o uvulare, ma mai postalveolare come in inglese)

dott.ssa CRISTINA MAZZALI ⟶ MAZZARI

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- **sviluppi possibili**

1	←	**B, P, F, V**	labiali, labiodentali
2	←	**C, K, G, J, Q**	velari
3	←	**S, Z, X**	sibilanti
4	←	**D, T**	dentali
5	←	**M, N**	nasali
6	←	**L, R**	liquide

X – aggiungere all'algoritmo la regola di semplificazione del nesso CS (del tipo 'clacson') in S

J – è un **problema**, sia come lettera iniziale sia nel corpo del nome: in nomi inglesi rappresenta la G palatale (John), in nomi francesi la prepalatale sonora (Julien), in italiano resta come semivocale nella grafia di nomi propri o di toponimi (Jacopo, Jesi)

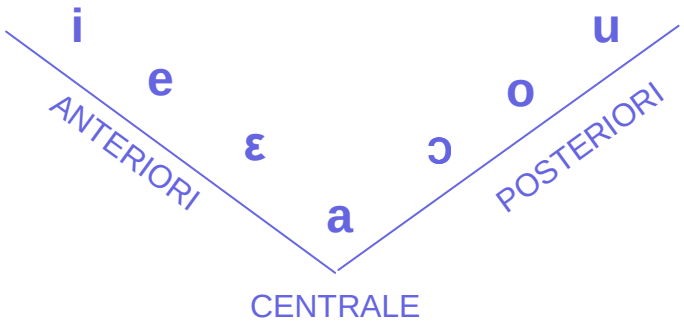
Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- **inglese e italiano**
- sviluppi possibili

Le differenze riguardano anche il trattamento delle **vocali**, che in italiano sono generalmente **ben pronunciate**.

Il sistema vocalico è, inoltre, piuttosto chiaro (sette vocali toniche, cinque atone) e crea difficilmente problemi di resa grafica (anche perché l'opposizione fonematica tra /e/ ed /ɛ/ e tra /ɔ/ e /o/ non ha riscontro grafematico), il che impone di valutare l'opportunità e il modo di accoglimento della regola della caduta delle vocali.



Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- **sviluppi possibili**

- A meno che non si lavori con codici fiscali, mantenere la **caduta delle vocali** e utilizzarle come **separatori**, in modo da conservare intatta la sostanza consonantica delle singole (e ben distinte) sillabe di un nome

  GERMANO → G06505 → G655

- Mantenere la **caduta di H**, considerato **non-separatore**, perché:
  - ✓ in it. è un segno grafico, non un grafema
  - ✓ si elimina il problema delle aspirate (es. Sahid)

- Mantenere la **caduta di W** (anche se nei nomi di origine tedesca non è una semiconsonante, ma suona come la labiodentale sonora /v/), **non-separatore** perché nei nomi di origine inglese è generalmente seguito da vocale

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- **sviluppi possibili**

Ulteriori affinamenti:

- Mantenere scempiamento delle consonanti geminate

  Nazzari → N360 / Nazari → N360

- Ridurre GL (+ I) a L, in modo da coincidere con LL (+ I) e L (+ I)

  Oglio / Ollio / Olio / [Olho] (resta probl. Oio < /ɔjo/) → O600

- Ridurre GN a N

  Ascanio, Eugenio; Mascagni, Montagna
  Bolaño, Robinho

- Ridurre SC (+ E, I) a S

  Si riducono i problemi dovuti a grafia incolta/distratta del digramma SC, ma si ottiene una stringa fuorviante nel caso in cui la lenizione della affricata prepalatale dia adito alla trascrizione di costrittiva prepalatale (Cesa, Pece, Aceto)

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- **sviluppi possibili**

Dovrebbero essere considerati, infine, anche i restanti **nessi consonantici**, allo scopo di individuare soluzioni efficaci in grado di ovviare agli errori di trascrizione dovuti a errata comprensione fonologica.

Si potrebbe pensare anche alla possibilità di:

- rivedere il criterio della lettera iniziale
  (→ codice numerico)
- sperimentare eventuali criteri di controllo
  (Reverse Soundex, conservazione prima vocale)

Problemi di Metodologia Statistica

3. La funzione 'Soundex':

- definizione
- utilizzo
- descrizione
- sviluppi e limiti
- altre funzioni
- sperimentazione
- inglese e italiano
- **sviluppi possibili**

Ciò appare, però, più pertinente a uno **studio futuro indipendente dall'algoritmo Soundex**, finalizzato allo messa a punto di una metodologia specifica per i legami tra record in lingua italiana.

Il passo successivo di questo studio-pilota sarà, invece, la **verifica sperimentale** – già in parte iniziata – delle soluzioni prospettate in questo incontro, allo scopo di accertare l'efficacia delle modifiche proposte all'algoritmo Soundex per il sistema linguistico italiano.