Similarità tra stringhe e poligonizzazione per uno studio diacronico dei toponimi

Ettore Sarzotti[1], Angelo Besana[1][0000-0002-1549-9591] Nicola Gabellieri[1][0000-0002-9220-9893], e Paolo Zatelli[2][0000-0003-3095-0472]

[1] Dipartimento di Lettere e Filosofia, Università di Trento, Via Tommaso Gar 14, 38122 Trento, Italy
[2] Dipartimento di Ingegneria Civile, Ambientale e Meccanica, Università di Trento, Via Mesiano 77, 38123 Trento, Italy

Abstract. Una fonte rilevante per la raccolta di toponimi storici è costituita dalle carte storiche. Questo tipo di documentazione pone il problema della digitalizzazione dei toponimi in ambiente GIS e del successivo confronto tra fonti diverse. In genere, questi processi vengono svolti manualmente e richiedono molto tempo. In questo lavoro si è cercato di automatizzare parzialmente il processo che porta alla costruzione di un geodataset di punti contenente la trascrizione dei toponimi per varie carte storiche di diversi periodi. Questo metodo è stato sviluppato in un caso di studio corrispondente a un comune della Provincia Autonoma di Trento (PAT), utilizzando una fonte attuale come la Carta Tecnica Provinciale (CTP) e tre fonti cartografiche storiche già considerate nella ricerca geografica (XIX-XX secolo). Inizialmente i toponimi sono stati raggruppati in base alla loro similarità basandosi sull'algoritmo Jaro-Winkler di distanza tra stringhe. Successivamente sono stati creati dei poligoni convessi dei punti appartenenti allo stesso gruppo per eseguire infine un join spaziale con i layer dei toponimi trascritti. In questo modo ogni toponimo è stato associato ai toponimi presenti nelle carte prese in esame, raggruppando toponimi affini e ottenendo una "biografia" di ogni toponimo. Il geodataset ottenuto si si pone come base per la realizzazione di ulteriori analisi. Il metodo semi-automatico presentato per la costruzione di un geodataset di toponomastica storica riduce significativamente il tempo di lavoro e il rischio di errore dell'operatore.

1 Introduzione

Da tempo la ricerca geografico-storica riconosce nella toponomastica una ricca fonte di informazioni per comprendere i cambiamenti paesaggistici e territoriali avvenuti nel corso del tempo nonché un valore di "patrimonio culturale immateriale" [1, 2]. Nel contesto degli studi diacronici, i toponimi possono servire come marcatori della continuità o dello spazio a varia scala. Vari studiosi hanno evidenziato come lo studio dei toponimi può rivelare le complesse relazioni tra le società umane e lo spazio [3–5].

La letteratura internazionale concorda sull'importanza di analizzare l'evoluzione di questi nomi tracciando la trasformazione dei territori, la persistenza della memoria culturale e l'impronta dell'attività umana [6, 7]. A questo proposito, una delle risorse per studiare l'evoluzione dei nomi dei luoghi è considerata la cartografia storica, che conserva una documentazione dei toponimi utilizzati nel passato. Una serie cartografica diacronica può restituire una "stratigrafia" di denominazioni spaziali utilizzate in diverse epoche, permettendo di registrare continuità e discontinuità e recuperare toponimi storici oggi scomparsi, come esperito in casi studio come la Toscana o il Trentino [8–10].
Nei fatti, la toponomastica rappresenta una "fonte nella fonte", ovvero metainformazioni che vengono estrapolate dalla documentazione cartografica [11]. Per questo motivo, la raccolta e l'analisi dei toponimi deve essere basata su metodi rigorosi e pone problemi interpretativi e tecnici. A questa criticità si vanno infatti a sommare le sfide in termini di georeferenziazione, digitalizzazione e analisi che pongono le carte storiche [10].
Da tempo sono stati sviluppati metodi per il corretto processamento delle carte storiche in ambiente GIS, che comprendono anche la raccolta della toponomastica in specifici geodataset.
I processi di estrazione e confronto dei toponimi dalle mappe storiche sono spesso basati sulla trascrizione e l'interpretazione manuale e risultano dispendiosi in termini di tempo [12]. Questo processo ad alta intensità di lavoro può aumentare il rischio di errore umano e limita la portata dell'analisi comparativa diacronica, soprattutto quando si ha a che fare con fonti cartografiche multiple e grandi quantità di dati.
Un esempio a questo proposito è costituito dal dataset che raccoglie la toponomastica storica raccolta da tre fonti cartografiche storiche per il territorio dell'attuale Provincia di Trento [13]. Tale operazione è stata compiuta manualmente per parte del territorio provinciale. La raccolta di elementi toponomastici da fonti diverse e il successivo confronto pone alcuni problemi relativamente alla localizzazione dell'elemento e al confronto con quelli contenuti nelle altre fonti cartografiche, che questo lavoro mira a affrontare.
Ponendosi in continuità con questo lavoro, il presente articolo espone un metodo semi-automatico per arricchire un set di dati geospaziali di toponimi storici con la "biografia" di ogni termine, cercando di affrontare e risolvere alcune delle criticità presentate. Tramite la creazione di uno script che utilizza un algoritmo di distanza tra stringhe i toponimi vengono raggruppati. Successivamente i punti appartenenti allo stesso gruppo vengono singolarmente raggruppati in cluster in base alla distanza, per poi creare Poligoni Convessi sulla base al gruppo e del cluster assegnati. Infine tramite join spaziali si arricchisce il layer poligonale con le ricorrenze dei toponimi appartenenti allo stesso gruppo del poligono, creando così una "biografia" del toponimo.
Questo approccio non solo riduce il tempo necessario per l'elaborazione dei dati, ma diminuisce anche il potenziale di errore dell'operatore, rendendolo uno strumento pratico per gli studiosi impegnati nella raccolta toponomastica da cartografia storica.

2 Fonti

Questo lavoro si propone di arricchire un processo di raccolta e digitalizzazione della toponomastica da fonti cartografiche storiche già esperito per il territorio della Provincia di Trento (Italia). L'obiettivo è quello di creare un geodataset in modo speditivo, ovviando ad alcune problematiche incontrate nel corso del lavoro quali la duplicazione di vari elementi toponomastici e l'esatta collocazione dell'elemento puntuale [12]. Per questa ricerca sono state utilizzate, oltre a una fonte odierna come la Carta Tecnica Provinciale (CTP), tre fonti cartografiche storiche zenitali disponibili per il territorio considerato: la *Carta d'Italia* dell'Istituto Geografico Militare (IGM) (1:25.000, 1927-31), il *Catasto fondiario austriaco* (1:2.880, 1853-1861) e il primo rilevamento militare asburgico (1:28.000, 1801-1805).

3 Caso Studio

Al fine di valutare protocollo metodologico semi-automatico, il metodo qua descritto è stato messo alla prova in un contesto territoriale limitato, coincidente con il Comune di Moena nella Provincia Autonoma di Trento. Il suo territorio comunale, situato nel comprensorio della Val di Fassa, è esteso circa 82,6 km^2. Per questa area è possibile contare su una lunga tradizione di studi toponomastici che offrono una significativa bibliografia di riferimento [13, 14]. La scelta del caso studio è motivata quindi dalla volontà di testare il metodo su un areale amministrativamente unitario, di sufficiente estensione, per il quale si ipotizza un cambiamento significativo dello strato toponomastico conseguente alla trasformazione socioeconomica e insediativa.

4 Metodo

Questo studio utilizza un processo semi-automatico per l'analisi dei toponimi storici, sfruttando i dati geospaziali preesistenti e la corrispondenza computazionale delle stringhe. La base di partenza è stato un dataset toponomastico che raccoglie come elementi puntuali toponimi delle quattro carte considerate.
Al fine di raggruppare i toponimi per similarità, è stato creato uno script in linguaggio R che è stato implementato in Qgis tramite il plugin "Processing R Provider" in uno strumento chiamato *Group Toponyms*. Prima di poter eseguire lo script è necessario creare manualmente un file .csv contenente i nomi comuni degli oggetti geografici quali ad esempio come "monte", "malga", "colle", che si vogliono escludere dal toponimo effettivo. Questo file viene dato come input allo strumento insieme al layer contenente i toponimi da raggruppare di tutte le carte prese in esame in un singolo campo della tabella attributi. In questo studio non sono stati considerati i toponimi di tipo lineare (in particolare relativi ai corsi d'acqua) che sono stati rimossi. Lo script, dopo aver eliminato i nomi degli oggetti geografici, confronta i toponimi e applica un algoritmo di somiglianza tra stringhe, che misura la somiglianza di due stringhe di caratteri basandosi su un concetto di distanza. L'algoritmo scelto è quello di Jaro-Winkler [15]

che esprime la distanza tra due stringhe con un numero compreso tra 0 e 1, dove 1 indica due stringhe identiche. Lo strumento permette di inserire come input un valore di soglia sopra il quale due stringhe sono considerate simili. Per questo studio è stato scelto empiricamente un valore soglia uguale a 0.9, ma il valore è ancora in fase di calibrazione e può essere adattato al caso specifico in esame. Sulla base di questa soglia lo script assegna ogni toponimo a un gruppo che viene immagazzinato in un nuovo campo della tabella attributi.
Dopodiché si devono compiere una serie di passaggi per poter creare, all'interno dei singoli gruppi, dei cluster basati sulla distanza in modo da poter tenere separati gruppi di toponimi che insistono su aree diverse. I passaggi sono i seguenti: 1. separare i gruppi di toponimi in singoli layer tramite il comando *Dividi vettore*; 2. creare dei cluster per ogni layer utilizzando lo strumento *DBSCAN clustering* in modalità *Processo in serie*; 3. riunire i vari layer con il comando *Fondi vettori* e creare un campo nuovo con un raggruppamento univoco derivato dalla combinazione del numero del gruppo di somiglianza e dell'ordine di cluster; 4. generare dei poligoni per ogni gruppo tramite lo strumento *Minima geometria di contorno* selezionando *Poligono Convesso* come tipo di geometria; 5. Creare un buffer attorno ai poligoni (in questo caso è stata scelta una distanza standard di 250 metri).
Successivamente si è proceduto a un join spaziale tra i poligoni con il buffer e il layer puntuale di partenza contenente tutti i toponimi. In tal modo la tabella attributi del layer poligonale è arricchita di quattro campi, ognuno con il toponimo presente in una specifica carta. Il join spaziale è del tipo "uno a molti" quindi si avranno riscontri multipli: per ogni poligono sono stati tenuti solamente i riscontri che appartengono allo stesso gruppo/cluster del poligono stesso. Questo processo ha garantito che ogni poligono contenesse la "biografia" dei toponimi associati, ovvero che ogni record riportasse la trascrizione dell'elemento toponomastico come registrata in ogni fonte. Questa struttura può facilitare ulteriori analisi diacroniche.
Infine a ogni gruppo è stato associato il toponimo appartenente alla carta più recente, con l'omissione dell'eventuale oggetto geografico, sostituendo così il raggruppamento numerico precedente, e per ogni poligono è stato poi ricavato un punto centroide che rappresenta il toponimo. Il centroide è stato ricavato anche per tutti i punti il cui gruppo è composto da soli due elementi e che quindi non hanno subito il processo di creazione dei poligoni convessi.

5 Risultati

Il processo di raggruppamento per similarità delle stringhe e il successivo clustering ha portato il numero di record da 345 del layer originale a 188. È possibile osservare questo cambiamento nella figura 1, dove sono raffigurati tutti i toponimi trascritti in origine, e in figura 2 dove sono rappresentati i toponimi che non hanno avuto raggruppamento insieme ai centroidi dei poligoni e dei gruppi composti da solo due elementi. Questo ha portato a una semplificazione nella lettura dei toponimi eliminando i punti con lo stesso toponimo all'interno di una stessa carta e raggruppandoli con toponimi affini di altre carte.

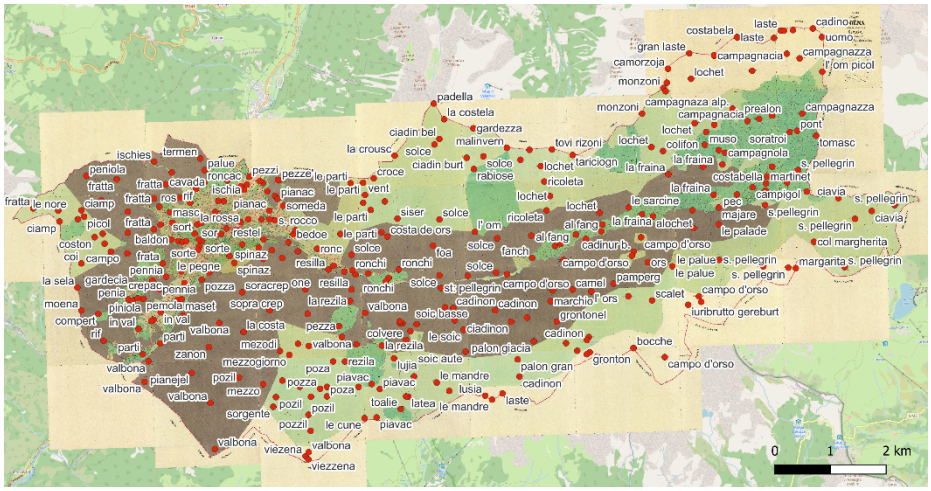

Fig. 1. Rappresentazione cartografica del layer contenente tutti i toponimi trascritti dalle quattro carte per il territorio del Comune di Moena. Sfondo: Catasto asburgico e Open Street Map.

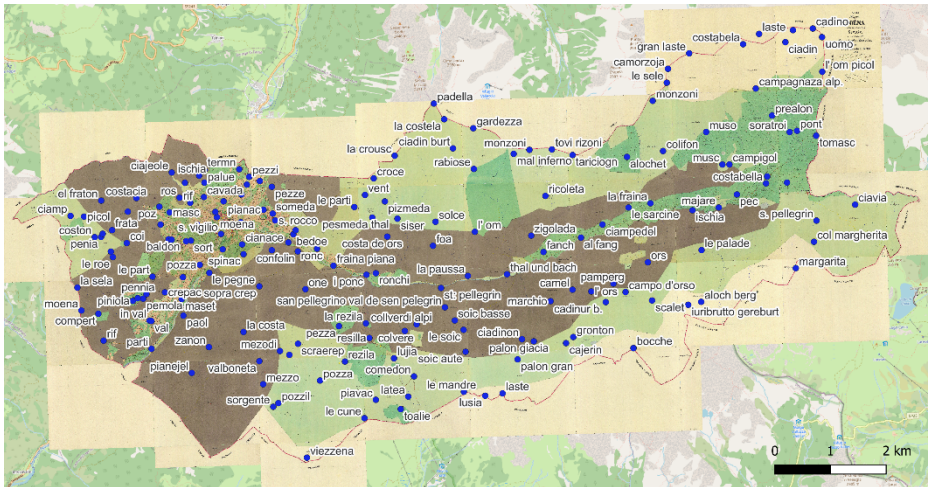

Fig. 2. Rappresentazione cartografica dei toponimi senza raggruppamento e i centroidi di poligoni e gruppi con soli due elementi per il territorio del Comune di Moena. Sfondo: Catasto asburgico e Open Street Map.

In figura 3 sono mostrati i poligoni convessi ottenuti dal raggruppamento dello script e dal processo di clustering. Inoltre in figura 4 si può osservare una porzione della tabella attributi dopo il join spaziale.

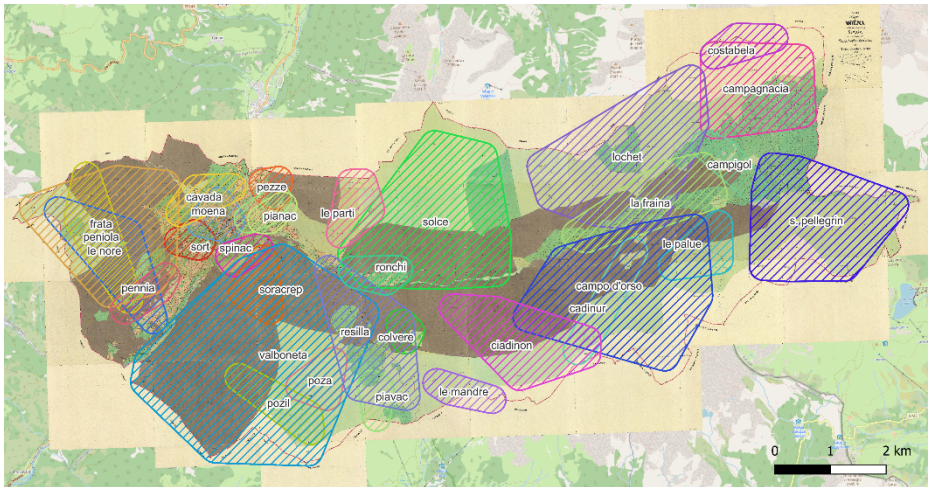

Fig. 3. Poligoni convessi dei gruppi di toponimi. Sfondo: Catasto asburgico e Open Street Map.

	group_cluster	name_group	CTP	IGM	Cat_Asb	C_fond_Austr
1	1-1	campagnacia	NULL	NULL	Campagnazza	NULL
2	1-1	campagnacia	NULL	NULL	Campagnazza	NULL
3	1-1	campagnacia	CIMA DE CAMPAGNACIA	NULL	NULL	NULL
4	1-1	campagnacia	STRADA DE CAMPAGNACIA	NULL	NULL	NULL
5	1-1	campagnacia	NULL	NULL	NULL	Campagnaza Alp.
6	11-1	pezze	NULL	NULL	Pezze'	NULL
7	11-1	pezze	NULL	Pezze	NULL	NULL
8	11-1	pezze	NULL	NULL	Pezze'	NULL
9	125-1	costabela	NULL	NULL	Cima di Costabella	NULL
10	125-1	costabela	CIMA DE COSTABELA	NULL	NULL	NULL
11	125-1	costabela	EL CIASTEL DE COSTABELA	NULL	NULL	NULL
12	147-1	campigol	SKILIFT CAMPIGOL	NULL	NULL	NULL
13	147-1	campigol	NULL	NULL	Campagnola	NULL
14	147-1	campigol	NULL	NULL	Lago di Campagnola	NULL
15	147-1	campigol	CAMPIGOL	NULL	NULL	NULL

Fig. 4. Tabella attributi del layer poligonale con i toponimi risultanti dal join spaziale raggruppati.

6 Discussione

Questo metodo rappresenta un primo tentativo di semplificare il layer dei toponimi raccolti da cartografia storica, aggregando toponimi “duplicati” perché estesi o perché presenti su due fogli diversi della stessa carta. Come metodo sperimentale presenta ancora alcune criticità. Al momento è stato applicato limitatamente al comune di Moena, per il quale il dataset originale è stato revisionato e corretto manualmente da alcuni errori, comprensibili vista la natura incline all’errore della trascrizione manuale

di un così vasto dataset. Ciò implica che, data la limitata estensione spaziale dello studio non si sono prodotti molti clustering all'interno dei gruppi. Inoltre lo script è solamente una versione iniziale e presenta ancora molti limiti come per esempio il fatto che i toponimi "*Resilla*" e "*Rezila*" non siano stati inseriti nel medesimo gruppo, oppure come l'algoritmo di Jaro-Winkler riconosca simile stringhe come "*Palue*" e "*Palue le*" ma totalmente diverse le stringhe "*Palue*" e "*Le Palue*". Nonostante queste iniziali problematiche il metodo è molto utile nel raggruppamento di toponimi e nel semplificare la lettura e l'analisi degli stessi.

7 Conclusioni

Il metodo descritto in questo studio presenta un approccio semi-automatico all'analisi toponomastica diacronica, migliorando in modo significativo l'efficienza e l'accuratezza dell'elaborazione dei nomi di luogo dalle mappe storiche. Partendo da un set di dati preesistenti di toponimi e sfruttando le misure di similarità delle stringhe attraverso la distanza di Jaro-Winkler, questo metodo consente di raggruppare sistematicamente i toponimi che probabilmente si riferiscono alla stessa caratteristica geografica. L'uso del clustering e dei poligoni convessi per definire le estensioni spaziali fornisce un mezzo strutturato e logico per ricostruire l'influenza territoriale di ciascun toponimo nel tempo.
Nonostante questo sia un primo tentativo di automatizzazione e presenti alcuni aspetti che possono essere ampiamente migliorati, l'innovazione chiave di questo approccio è l'integrazione di moderni strumenti computazionali, in particolare l'automazione della corrispondenza delle stringhe e delle giunzioni spaziali, per snellire quello che è un processo manuale che richiede molto tempo. Automatizzando il raggruppamento dei toponimi e arricchendo i poligoni risultanti con dati storici, questo metodo consente di creare rapidamente un geodataset sincronico e diacronico dei nomi dei luoghi in diversi periodi.
Applicata al caso di studio del comune di Moena, questa metodologia ha dimostrato il suo potenziale nel ridurre l'errore dell'operatore e nell'accelerare l'analisi di fonti storiche complesse. Sebbene la trascrizione dei toponimi richieda ancora uno sforzo manuale, l'automazione delle fasi successive riduce notevolmente il tempo necessario per l'analisi. Questo metodo è particolarmente prezioso per gli studi su larga scala in cui il confronto manuale dei toponimi sarebbe proibitivo in termini di tempo.
Nelle ricerche future, questo approccio potrebbe essere esteso ad altre regioni o insiemi di dati, potenzialmente incorporando ulteriori mappe storiche o utilizzando soglie diverse per la somiglianza delle stringhe per affinare i risultati. In definitiva, questo metodo fornisce una utile strategia metodologica per i ricercatori interessati alla raccolta di toponomi storici e alla costruzione di Historical GIS toponomastici.

Riconoscimenti. Il presente lavoro è finanziato dall'Unione europea – Next Generation EU, nell'ambito del bando PRIN 2022, progetto *"Bridging geography and history of woodlands: analysing mountain wooded landscapes through multiple sources and historical GIS"* (2022EKECST) – CUP E53D23010170006.

Riferimenti bibliografici

1. Cantile, A., Kerfoot, H. a c di: Place names as intangible cultural heritage. IGMI, Firenze (2016).
2. Gelling, M.: Signposts to the past: place-names and the history of England. edn. Dent, London (1979).
3. Cassi, L., Marcaccini, P.: Toponomastica, beni culturali e ambientali. Gli «Indicatori geografici» per un loro censimento. Società Geografica Italiana, Roma (1998).
4. Cassi, L.: From Historical to New Place Names. The Case of Italy. In: O'Reilly, G. (a cura di) Place Naming, Identities and Geography. pp. 575–599. Springer International Publishing, Cham (2023).
5. Dematteis, G.: Le metafore della terra: la geografia umana tra mito e scienza. edn. Feltrinelli, Milano (1985).
6. Sousa, A., García-Murillo, P.: Can place names be used as indicators of landscape changes? Application to the Doñana Natural Park (Spain). Landscape Ecology. **16** (5), 391–406 (2001).
7. Penko Seidl, N.: Engraved in the Landscape: The Study of Spatial and Temporal Characteristics of Field Names in the Changing Landscape. Names. **67** (1), 16–29 (2019).
8. Gabellieri, N., Grava, M.: A changing identity: from an agrarian and manufacturing region to a multi-functional territory. In: Cantile, A., Kerfoot, H. (a cura di) Place names as intangible cultural heritage. pp. 140–160. IGMI, Firenze (2016).
9. Dai Prà, E., Gabellieri, N., Scanu, N.: In the footsteps of Cesare Battisti: an approach for collection and analysis of toponyms using historical maps and HGIS in Trentino, Italy (19th-21st centuries). In: Cantile, A., Kerfoot, H. (a cura di) Permanence, transformation, substitution and oblivion of geographical names 3rd International Scientific Symposium (Castel dell'Ovo, Naples, Italy, 22th - 24th September 2021). pp. 41–50. IGMI, Firenze (2022).
10. Grava, M., Berti, C., Gabellieri, N.: Historical GIS: strumenti digitali per la geografia storica in Italia. EUT, Edizioni Università Trieste, Trieste (2020).
11. Fuchs, S.: Toponymic GIS. Role and potential of place names in the context of geographic information systems and GIS. KN - Journal of Cartography and Geographic Information. **65** (6), 330–337 (2015).
12. Panecki, T.: Mapping Imprecision: How to Geocode Data from Inaccurate Historic Maps. ISPRS International Journal of Geo-Information. **12** (4), 149 (2023).
13. Dai Prà, E., Gabellieri, N., Scanu, N.: Dalla mappa al geodatabase: un modello di raccolta, digitalizzazione e analisi sincronica e diacronica in ambiente GIS del patrimonio toponomastico del territorio trentino da fonti cartografiche storiche (XIX-XXI secolo). Il capitale culturale. Studies on the Value of Cultural Heritage. (25), 603–633 (2022).
14. Cordin, P., Flöss, L., Gatti, T.: Il Dizionario toponomastico trentino-DTT: dalla ricerca geografica alla ricerca storica. Studi trentini di scienze storiche. **90**, 469–496 (2011).
15. Wang, Y., Qin, J., Wang, W.: Efficient Approximate Entity Matching Using Jaro-Winkler Distance. In: Bouguettaya, A., Gao, Y., Klimenko, A., Chen, L., Zhang, X., Dzerzhinskiy, F., Jia, W., Klimenko, S.V., Li, Q. (a cura di) Web Information Systems Engineering – WISE 2017. pp. 231–239. Springer International Publishing, Cham (2017).