ZfdG

Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Kontextsensitive Entscheidungsfindung zur automatisierten Identifizierung und Clusterung deutschsprachiger Urbanonyme

Autor/in:
Jan Michael Goldberg

Kontakt:
jan.goldberg@wiwi.uni-halle.de

Institution:
Martin-Luther-Universität Halle Wittenberg, Lehrstuhl für empirische Makroökonomik

GND:
1240406630

ORCID:
0000-0002-4817-4283

DOI des Artikels:
10.17175/2022_005

Nachweis im OPAC der Herzog August Bibliothek:
1817560271

Erstveröffentlichung:
10.10.2022

Letzte Überprüfung aller Verweise:
03.08.2022

GND-Verschlagwortung:
Historische Geografie | Cluster-Analyse | Lokalisation | Kontextbezogenes System | Klassifikation |

Zitierweise:
Jan Michael Goldberg: Kontextsensitive Entscheidungsfindung zur automatisierten Identifizierung und Clusterung deutschsprachiger Urbanonyme. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel 2022. PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/2022_005.

Jan Michael Goldberg

Kontextsensitive Entscheidungsfindung zur automatisierten Identifizierung und Clusterung deutschsprachiger Urbanonyme

Abstracts

Viele historische Quellen enthalten zahlreiche Ortsangaben, deren manuelle Zuordnung viele Ressourcen bindet. Um hier Abhilfe zu schaffen wird ein Algorithmus beschrieben, mit dem solche Urbanonyme automatisiert geokodiert werden können. Ebenso ist es möglich, die Orte entsprechend ihrer gemeinsamen historischen Verwaltungszugehörigkeit zu clustern. Probleme wie gleiche Namen bei Ortsbezeichnungen werden vor allem durch eine Einbeziehung weiterer Informationen desselben Kontextes (derselben Quelle) gelöst. Eine Validierung geschieht anhand von etwa 3,4 Millionen überwiegend deutschsprachigen Ortsangaben aus der genealogischen Datenbank *GEDBAS*. Zusammenfassend lassen sich etwa drei von vier relevanten Ortsangaben identifizieren und lokalisieren. Über 90 Prozent der identifizierten Ortsangaben können ihrer historischen Provinz zugeordnet werden.

Many historical sources contain numerous names of places, the manual assignment of which ties up a lot of resources. To simpliy this, an algorithm is described with which such urbanonyms can be geocoded automatically. It is also possible to cluster the places according to their common historical administrative affiliation. Problems such as identical terms for place names are solved primarily by including further information from the same context (same source). A validation is done on the basis of about 3.4 million mostly German-language place names from the genealogical database *GEDBAS*. In summary, about three out of four relevant place names can be identified and localised. More than 90 percent of the identified place names can be assigned to their historical province.

1. Einleitung

Orte werden tagtäglich in vielen Applikationen identifiziert und lokalisiert.[1] Der Bahnticketautomat, das Navigationsgerät im Auto oder die Suchmaschine im Internet stellen dafür nur einige willkürliche Beispiele dar. Das Prinzip dahinter ist vielfach aber gleich: Nach der Eingabe eines Orts oder mindestens eines Teils davon schlägt die Applikation einen oder mehrere auszuwählende(n) Orte vor. Die letztendliche Auswahl kann vom Menschen getroffen werden. Steht jedoch kein Mensch zur Verfügung, um eine abschließende Auswahl zu treffen, muss die Entscheidung auf zuvor definierten Kriterien beruhen. Durch Dopplungen, Ähnlichkeiten und Variationen von Ortsnamen ist dies jedoch kein triviales Unterfangen.

In Deutschland (und Europa) gibt es beispielsweise zahlreiche Orte namens Neustadt. Allein 36 davon sind in der Arbeitsgemeinschaft Neustadt in Europa zusammengefasst.[2] Hinzu kommt, dass etliche Stadtteile diese Bezeichnung tragen. Werden die historischen, heute nicht mehr genutzten Bezeichnungen inkludiert, sind es nochmals mehr. Das kann zum einen darin begründet sein, dass die Ortsbezeichnungen heute einfach nicht mehr in Verwendung sind, weil es den Ort entweder nicht mehr gibt oder aber eine Umbenennung stattgefunden hat.[3] Zum anderen können die historischen Ortsangaben aber auch einen Ort beschreiben, dessen Relevanz heute nicht mehr ausreichend ist, um diesen durch eine gesonderte Bezeichnung zu würdigen – ohne, dass dieser gänzlich verschwunden wäre. Möglicherweise ist der Ort auch in einem übergeordneten aufgegangen; kleine Siedlungsformen, wie beispielsweise Weiler, verschwanden im Laufe der letzten Jahrhunderte. Die Herausforderungen, die sich heute aus der Lokalisierung von Ortsnamen ergeben, sind bei der Zuordnung historischer Bezeichnungen – im Weiteren als Urbanonyme[4] bezeichnet – also umso größer.

Historische Ortsangaben stehen allerdings selten für sich allein, sondern sind von Kontextinformationen umgeben. Eingebettet in einen Fließtext kann sich aus den weiteren Informationen ergeben, um welchen Ort es sich handelt. Beispielsweise können in einem Adressbuch andere Ortsangaben oder die Angabe einer übergeordneten Gebietskörperschaft vorhanden sein. Diese als Kontext bezeichneten Informationen können zur Identifizierung genutzt werden. Nachfolgend kann auf dieser Basis eine geographische Lokalisierung des Orts angestellt werden. Im Kontext dieser Arbeit meint Lokalisierung eine Ortsbestimmung, die Objekte zu einer Adresse im physischen Raum der Erdoberfläche zuordnet.[5] Eine gesonderte, automatisierte Lösung

[1] Die Identifizierung beschreibt die Zuordnung zu einem Identifikator, der einem physischen Ort zugeordnet ist. Dagegen umfasst die Lokalisierung die gelungene geographische Verortung des identifizierten Ortes.
[2] Working Group Neustadt in Europa (Hg.) 2021.
[3] Zedlitz / Luttenberger 2014, S. 218–231.
[4] Der Begriff der Urbanonyme beschreibt dabei Siedlungsstrukturen (z. B. Städte, Dörfer) und stellt somit eine Unterkategorie der Toponyme da. Er wird im Folgenden synonym zum Begriff der Ortsbezeichnung verwendet.
[5] Die Bestimmung der Position wird heute oftmals satellitengestützt realisiert (vgl. Gentile et al. (Hg.) 2013). Diese Möglichkeit setzt voraus, dass das zu lokalisierende Objekt (1.) bekannt ist und sich (2.) an der zu lokalisierenden Position befindet. Historische Ortsangaben hingegen beschreiben Positionen von (unbekannten) Objekten zu vergangenen Zeitpunkten. Deshalb sind moderne Möglichkeiten der Lokalisierung auf diese Problematik nicht anwendbar.

zur Identifikation und Lokalisierung (speziell deutschsprachiger) historischer Urbanonyme ist bis dato nicht bekannt. Diese Lücke wird durch den vorliegenden Artikel geschlossen, indem ein Algorithmus als Lösung vorgeschlagen wird, der historische Ortsangaben[6] kontextsensitiv lokalisiert. Eine Übersicht der Begriffliche ist in Abbildung 1 vorhanden.

Für viele wissenschaftliche Fragestellungen reicht allein die Lokalisierung der Ortsangaben jedoch nicht aus. Vielmehr ist es auch wichtig, die (historische) administrative Zugehörigkeit der Orte zu ermitteln und alle zusammengehörigen Ortsangaben zu clustern. Darum wird mit dem Algorithmus auch eine Methode bereitgestellt, mit der Orte über die administrative Zugehörigkeit zu einem definierten Zeitpunkt geclustert werden können. Der Algorithmus wird primär mit Hilfe von Urbanonymen aus genealogischen GEDCOM-Dateien[7] aus der Datenbank Genealogische Datenbasis (GEDBAS) getestet, soll aber auch auf andere Quellen adaptiert werden können. Damit wird ebenfalls einem Vorschlag Gellatlys nachgekommen, eine Software zur Geokodierung von Ortsangaben zu entwickeln.[8]

Im folgenden Kapitel werden zunächst verschiedene bestehende Lösungsansätze betrachtet, auf deren Basis die Entwicklung des Algorithmus stattfindet. Danach wird der entwickelte Algorithmus in der Programmiersprache Python umgesetzt und am Beispiel von GEDCOM-Dateien aus der GEDBAS validiert. Abschließend erfolgt eine Zusammenfassung.

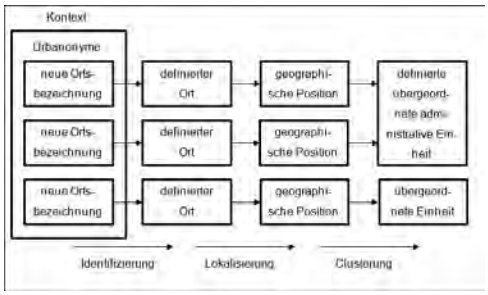

Abb. 1: Übersicht über Begrifflichkeiten und Zusammenhänge. [Goldberg 2022]

2. Identifizierung und Lokalisierung von Urbanonymen

Die Identifizierung beschreibt die Zuordnung eines Urbanonyms (z. B. ›Berlin‹) zu einer physisch existierenden Entität – einem Ort (z. B. der Hauptstadt Berlin der Bundesrepublik Deutschland). Die Lokalisierung hingegen besteht in der Zuordnung von Koordinaten zu diesem Ort. Auch wenn die Lokalisierung das wesentliche Ziel darstellt, so liegt in der vorhergehenden Identifizierung die maßgebliche Herausforderung. Das ist dadurch bedingt, dass zu nahezu allen (identifizierten) Orten die geographischen Koordinaten leicht zu ermitteln sind, die eindeutige Identifizierung jedoch durch verschiedene historisch bedingte Faktoren erschwert wird.

Im Folgenden wird der Stand der Technik verschiedener Teilaspekte dieser Herausforderung beschrieben. Zunächst wird erläutert, wie der Kontext zur Entscheidungsfindung beitragen kann. Danach werden kurz relevante Suchalgorithmen sowie Methoden der Ähnlichkeitsanalyse aufgegriffen. Da eine Identifizierung von Urbanonymen nur unter Hinzunahme (historischer) Ortsverzeichnisse möglich ist, finden auch diese Beachtung. Abschließend wird in die Struktur von GEDCOM-Daten eingeführt, die in dieser Studie zur Validierung des Algorithmus dienen.

2.1 Kontextsensitive Entscheidungsfindung

Das oben angeführte Beispiel Berlins zeigt die Problematik deutlich auf: Neben der Hauptstadt könnte ebenso eine andere Entität, z. B. das Land Berlin der Bundesrepublik Deutschland oder der Ortsteil Berlin der schleswig-holsteinischen Gemeinde Seedorf, gemeint sein.[9] Ohne weitere Informationen, worauf sich die Ortsangabe bezieht oder in welchem Zusammenhang diese genannt wird, kann eine Identifizierung nicht eindeutig vorgenommen werden. Die Informationen im Zusammenhang der Verwendung eines Wortes werden dabei als ›Kontext‹ bezeichnet. Nach Dey und Abowd ist Kontext aus informationstechnischer

[6] Als *historische Ortsangaben* werden im Folgenden solche Bezeichnungen beschrieben, die in deutscher Sprache einen Ort im deutschsprachigen Raum im Zeitfenster der Neuzeit beschreiben. In Abgrenzung dazu finden Ortsbezeichnungen aus antiken und mittelalterlichen Quellen keine Verwendung.
[7] Sie enthalten oftmals sehr viele Ortsangaben zu Lebensereignissen wie Geburten oder Hochzeiten. GEDCOM-Dateien stellen dabei den gängigen Standard im Austausch genealogischer Informationen dar (vgl. Gellatly 2015, S. 112; Harviainen / Björk 2018).
[8] Gellatly 2015, S. 118.
[9] In dem Fall würde es sich bei der Ortsangabe um ein Choronym, nicht um eine Urbanonym, handeln.

Sicht jede Information, die zur Charakterisierung der Situation einer Entität genutzt werden kann.[10] Systeme, die diesen Kontext nutzen, werden als ›kontextsensitiv‹ bezeichnet.[11] Kontextsensitive Systeme bedingen also eine Flexibilität in der Ausführung eines Prozesses. Der Kontext als externer Einfluss kann dazu führen, dass der interne Prozess der Informationsverarbeitung angepasst wird.[12] Ein Beispiel dafür ist die Anzeige der Wettervorhersage in Abhängigkeit der Position oder die Veränderung der Bildschirmhelligkeit in Abhängigkeit der Umgebungsbeleuchtung. Auch in CARS (Context Based Recommendation Systems) werden Nutzer*innen auf Basis der zu ihnen verfügbaren Kontextinformationen in der Entscheidungsfindung unterstützt.[13] Daneben ist eine eigenständige Entscheidung zwischen mehreren konkreten Alternativen ohne menschliches Zutun auf Basis des Kontextes möglich.

Bei der Lokalisierung ist eine Entscheidung notwendig, um die Verbindung zwischen einer konkreten Ortsbezeichnung und einem Ort zu definieren. Solche Entscheidungen können (in Anlehnung an einen binären Klassifikator) richtig oder falsch sein. Jedoch kann auch keine Entscheidung getroffen werden – und auch dieses kann richtig oder falsch sein. Richtig kann eine nicht getroffene Entscheidung für einen Ort beispielsweise sein, wenn die Ortsbezeichnung ›John Michael‹ lautet und kein Urbanonym darstellt, sondern – wie im Beispiel – womöglich aus einem Eingabefehler resultiert und einen Vornamen darstellt (TN). Diese Konstellationen sind in Tabelle 1 dargestellt. Dasselbe Schema ist auf die Lokalisierung und die nachfolgende regionale Klassifizierung anwendbar. Ziel ist es, möglich viele T*-Zuordnungen zu erreichen, gleichzeitig aber die F*-Rate niedrig zu halten.

	Identifizierung korrekt	Identifizierung nicht korrekt
Identifizierung erfolgt	True positive (TP)	False positive (FP)
Identifizierung nicht erfolgt	True negative (TN)	False negative (FN)

Tab. 1: Konfusionsmatrix zur Identifizierung von Ortsbezeichnungen in Anlehnung an Fawcett. [Fawcett 2006, S. 862]

Um es am Beispiel Berlins auszudrücken: So würde in einer Liste der Bundesländer eher die Entität des Landes aufgegriffen, in einer Liste aller Hauptstädte jedoch die Entität Berlins als Hauptstadt Deutschlands. Der Kontext würde hier den Schluss zulassen, ob es sich um das Land oder die Stadt handelt. Diese Schlussfolgerung ist das Ergebnis einer Heuristik: Alle Werte in der Liste beschreiben Länder, also ist es wahrscheinlich, dass der letzte Wert auch ein Land ist. Jedoch kann der Wert auch kein Land darstellen, wodurch die Schlussfolgerung falsch wäre (FP). Heuristiken garantieren keine richtigen Ergebnisse,[14] sie dienen der Findung einer wahrscheinlich korrekten Lösung unter begrenztem Wissen und wenig Zeit.[15] Insofern sind kontextsensitive Entscheidungen, wenn der Kontext die Entscheidung nicht eineindeutig zulässt, heuristische Verfahren. Um eine Heuristik im Entscheidungsprozess eines technischen Systems einzusetzen, ist ihre Formalisierung notwendig. Eine programmtechnische Formalisierung der Heuristik führt in der Folge zu einem (heuristischen) Algorithmus.

Um den Kontext einer Ortsbezeichnung nun für die Zuordnung zu einer Entität zu verwenden, müssen die Heuristiken definiert werden, die die Entscheidungsfindung beeinflussen. Zandhuis et al. nennen drei Möglichkeiten der Einbindung des Kontextes:

1. den Ort, an dem die Ortsangabe erstellt wurde,
2. eine Gebietszugehörigkeit und
3. die Zeit, in der die Quelle kreiert wurde in Zusammenhang mit der Information, wann ein Ort wie bezeichnet wurde, da die Bezeichnung temporalen Schwankungen unterliegen kann.[16]

Insbesondere bei Sekundärquellen ist der Ort der (Primär-)Quellenerstellung nicht relevant oder bekannt. Auch Gebietsangaben sind nicht immer vorhanden. Ebenso verhält es sich mit den temporalen Informationen. Zudem kann in Sekundärquellen eine sprachliche Anpassung des Ortsnamens (an die Schreibweise zur Zeit der Erstellung der Sekundärquelle) stattgefunden haben.

Aufgrund dieser Schwierigkeiten sind die drei genannten Punkte für eine Identifizierung nicht ausreichend. Es lassen sich jedoch andere Kontextinformationen finden: Der Kontext von Ortsbezeichnungen kann aus weiteren Ortsangaben bestehen, die möglicherweise in einem Zusammenhang stehen. Gegebenenfalls sind im Kontext auch konkrete Angaben zur geographischen Distanz zwischen den Entitäten, zur gemeinsamen administrativen Zugehörigkeit oder einer sonstigen Beziehung vorhanden. Zudem können temporale Angaben im Kontext die Identifizierung der Ortsangaben unterstützen. Aber nicht nur die Nennung von Orten, sondern auch die von Namen kann relevant sein: Die Häufigkeit von Vornamen variiert über die Zeit, sodass hierüber

[10] Abowd / Dey 1999, S. 1, 3-4.
[11] Sitou 2009, S. 28, 123.
[12] Rosemann / Recker 2006, S. 149.
[13] Zheng et al. 2019, S. 2453.
[14] Feigenbaum / Feldman (Hg.) 1963, S. 6.
[15] Vgl. Gigerenzer / Todd (Hg.) 1999, S. 147.
[16] Zandhuis et al. 2015, S. 36.

eine Schätzung des Geburtsjahres – und somit eine zeitliche Verortung der Ortsbezeichnung – vorgenommen werden kann.[17] Wichtiger jedoch erscheint die regionale Dichte und Häufigkeit von Namen: Die Verwendung von Nachnamen ist oftmals geografisch nicht gleichverteilt.[18] In Abbildung 2 sind beispielhaft die relative und absolute Verteilung des Nachnamens ›Hinse‹ abgebildet. Es ist eine erhöhte Dichte im westfälischen Raum zu erkennen.[19] Demzufolge ist also wahrscheinlicher, dass der Name in Kombination mit Orten im westfälischen Raum genannt wird.

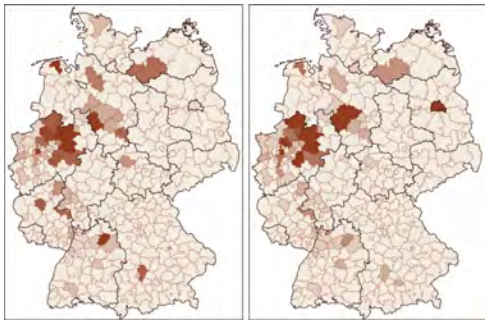

Abb. 2: Relative (links) und absolute (rechts) Verteilung des Nachnamens Hinse. [Goldberg 2022, erstellt mit Geogen Deutschland, Stöpel 2021a.]

Auch bei der Vornamensgebung sind in Deutschland regionale Unterschiede zu erkennen,[20] was teilweise auf unterschiedliche Präferenzen in der Namensgebung verschiedener Konfessionen zurückzuführen ist.[21]

2.2 Suchalgorithmen und Ähnlichkeitsanalyse

Alle (vormals) existierenden Orte können in einer Liste dargestellt werden. Sie stellen eine endliche Menge dar. Da es jedoch hunderttausende Orte gibt (oder gab), von denen jeweils einer ausgewählt werden soll, ist anzunehmen, dass die Auswahl eines Suchalgorithmus erhebliche Auswirkung auf die Performance hat. Zur Suche in Listen existieren verschiedenste Algorithmen, die oftmals an ihre jeweilige Anwendung angepasst werden. Da es sich bei Ortsbezeichnungen um Zeichenketten handelt, sind hier insbesondere String-Matching-Algorithmen relevant. Die Namen der Orte fungieren dabei als Eigenschaft der Entität, die mit der Ortsbezeichnung verglichen werden kann.

Suchalgorithmen können grundlegend in einfache und informierte Suchen unterteilt werden. Im Gegensatz zu den einfachen Suchen ist bei den informierten Suchen ein Wissen über den Suchraum vorhanden.[22] Bei einer gegebenen Liste von Orten findet ein einfaches Suchverfahren Verwendung, bei der Auswahl eines Suchalgorithmus sind dagegen insbesondere Laufzeit und Genauigkeit von Interesse: Während bei einer linearen Suche in Listen jedes Element betrachtet werden muss, ist bei der binären Suche nur die logarithmische Anzahl von Vergleichen durchzuführen. Es zeigt sich, dass bei der Verarbeitung vieler Urbanonyme soweit wie möglich auf lineare Suchen verzichtet werden sollte; stattdessen ist der Einsatz binärer Suchalgorithmen angebracht. Hierzu ist vorbereitend eine alphabetische Sortierung möglicher Zielorte durchzuführen.

Ortsbezeichnungen können jedoch Rechtschreibfehler wie vertauschte Zeichen aufweisen. Ebenso kann die Schreibweise von Orten im Zeitverlauf einer Variation unterliegen, ohne dass eine gänzliche Umbenennung erfolgt ist. Auch wenn die Suche keine eindeutigen Treffer hervorbringt, kann über einen Vergleich der Ähnlichkeit eine Identifizierung erfolgen. Zum Vergleich der Ähnlichkeit gibt es verschiedene Maße. Ein übliches Maß stellt die Levenshtein-Distanz dar, eine Metrik, mit der Schritte der Veränderung zwischen Wörtern gezählt werden.[23] Die Distanz ist bei identischen Zeichenketten 0 und beträgt maximal die Länge des längeren Strings. Hier besteht die Herausforderung, einen Wert auszuwählen, der für den jeweiligen Vergleich als plausibel gilt,[24] wobei ein Bezug auf die Länge der gesuchten Zeichenkette angebracht sein kann.[25]

[17] Gallagher / Chen 2008.
[18] Vgl. Seibicke 1982, S. 148.
[19] Die Karte wurde mit Geogen erzeugt und basiert auf etwa 35 Millionen Telefonbucheinträgen. Stöpel 2021b.
[20] Gesellschaft für deutsche Sprache e. V. (Hg.) 2019.
[21] Vgl. Seibicke 1982, S. 150.
[22] Vgl. Clarke et al. 2009, S. 34.
[23] Levenštejn 1966.
[24] Es sind keine allgemein üblichen Toleranzkriterien bei dem Vergleich von Ortsstrings bekannt. Möglicherweise ist eine solche Art der Ähnlichkeitsanalyse bei Orten nicht besonders zielführend, da viele Orte sehr gleichklingend sind und eine Levenshtein-Distanz von 2 bereits wesentliche Veränderungen herbeiführen kann.
[25] Vgl. List 2010, S. 8.

Statt in einer Liste können Orte auch in Bäumen strukturiert sein. Zur Optimierung des Suchverhaltens in diesen ist ein informierter Ansatz möglich, der beispielsweise nur solche Pfade mit definierten Eigenschaften durchsucht.

2.3 Historische Orte und Urbanonyme

Um überhaupt Orte zu haben, die als Entitäten dienen und einem Vergleich mit der Ortsbezeichnung zugefügt sind, ist die Auswahl eines Ortsverzeichnisses eine notwendige Bedingung. Im Weiteren wird dazu das Geschichtliche Orts-Verzeichnis (GOV) des Vereins für Computergenealogie Verwendung finden.[26] Dieses enthält etwa 1,2 Millionen Objekte[27] in Europa.[28] Das GOV wird ausgewählt, weil es besonders im deutschsprachigen Raum viele historische Urbanonyme abbildet. Für andere Länder, beispielsweise die Niederlanden, gibt es zudem weitere ausführliche Portale.[29]

Zugang zum GOV kann eine externe Anwendung auf zwei Arten erlangen. Zum einen über das sogenannte Mini-GOV worin verschiedene Informationen zu Orten enthalten sind: Als *Uniform Resource Identifier* (URI) eines Ortes dient die sogenannte GOV-Kennung. Daneben ist der Typ des Objekts vorhanden, gefolgt von dem aktuellen Namen (im Falle von früheren deutschen Siedlungsgebieten auch der letzte deutsche Name). Auch der Staat, dem das Objekt angehört, sowie vier Angaben zur administrativen Zuordnung werden angegeben.[30] Zudem ist die Postleitzahl vorhanden. Abschließend folgen mit der Längen- und Breitengrade die Koordinaten des Objekts. Zum anderen besteht die Möglichkeit über einen Webservice auf die Daten im GOV zuzugreifen. Während durch das Mini-GOV nur die oben genannten Informationen zu einem Ort zur Verfügung stehen, kann über den Webservice auf sämtliche Daten des GOV zugegriffen werden (siehe Abbildung 3).[31] Die Abfrage des Webservice ergibt ein assoziatives Datenfeld (im Folgenden Dictionary), in dem verschiedene Informationen zum Ort enthalten sind. Die Beschreibung der Zugehörigkeit zu übergeordneten Objekten erfolgt im Key ›part-of‹. Das in Abbildung 3 dargestellte Beispiel enthält vier Zugehörigkeiten. Die Dauer der Zugehörigkeit kann entweder über ›timespan‹ definiert sein. Im anderen Fall sind – wie im Beispiel – Beginn und Ende gesondert definiert. Auch ist die GOV-ID des übergeordneten Objekts enthalten. Das ermöglicht beispielsweise zusätzlich die administrativen Zusammenhänge der Vergangenheit zu erfassen. Die dadurch entstehende Baumstruktur ist implizit, da sie erst während der Suche erzeugt wird.

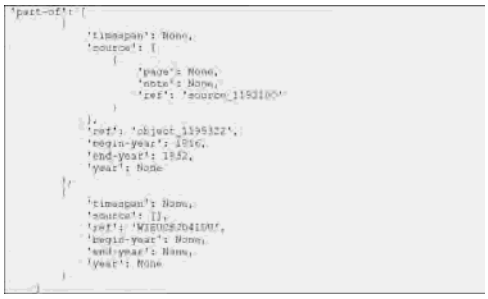

Abb. 3: Auszug aus einer GOV-Abfrage. [Goldberg 2022]

Urbanonyme kommen in vielen Quellen vor. Besonders konzentriert sind Urbanonyme in genealogischen Datenstrukturen zu finden. Hierin sind es die zentralen Vitaldaten von Personen (Geburt bzw. Taufe, Heirat und Tod bzw. Beerdigung), die oftmals auch mit einer Ortsbezeichnung versehen sind. Durch die verwandtschaftlichen Verknüpfungen liegen zudem verschiedene Kontextdaten vor (Namen, Zeiten, weitere Ortsangaben). Aufgrund der hohen Informationsdichte werden im Zuge dieser Arbeit genealogische Daten in Form von GEDCOM-Dateien zur Validierung des Algorithmus eingesetzt. Einzelne GEDCOM-Dateien bestehen aus Text mit einer definierten Syntax. Eine Person wird beispielsweise wie im nachfolgenden Beispiel definiert (siehe Abbildung 4). Verschiedenartige Informationen werden dabei mit verschiedenen Tags gekennzeichnet. Standardmäßig

[26] Das GOV ist nicht das einzige Verzeichnis, dass historische Ortsangaben zusammengeführt. Daneben gibt es verschiedene Ortslexika. Ein online-verfügbares Beispiel stellt das Historische Ortsverzeichnis von Sachsen dar (vgl. Institut für sächsische Geschichte und Volkskunde (Hg.) 2020). Für den gesamten deutschsprachigen Raum ist das GOV jedoch die einzige digitale, klar strukturierte Sammlung historischer Ortsangaben. Zudem unterliegt das Mini-GOV einer Creative Common Lizenz und kann deshalb in diesem Rahmen Verwendung finden.
[27] Diese Objekte sind jeweils sogenannten Typen zugeordnet. Die Typen beschreiben nicht nur Urbanonyme, sondern auch politische, kirchliche oder gerichtliche Verwaltungen; auch geographische Objekte oder solche zum Verkehrswesen sind vorhanden.
[28] Verein für Computergenealogie e. V. (Hg.) 2015. Das GOV erscheint insbesondere für den deutschsprachigen Raum geeignet.
[29] Vgl. Zandhuis et al. 2015, S. 24.
[30] Die administrative Zuordnung unterscheidet sich je nach heutigem Staat. In Deutschland ist die adm. Zuordnung 1 auf das Bundesland bezogen, die adm. Zuordnung 2 auf den Regierungsbezirk, die adm. Zuordnung 3 auf den Kreis, Landkreis, Stadtkreis die kreisfreie Stadt oder die Region und die adm. Zuordnung 4 auf die Stadt, Gemeinde, den Markt oder Flecken.
[31] Verein für Computergenealogie e. V. (Hg.) 2021a.

definiert der Tag PLAC PLAC dabei eine Ortsangabe.[32] Diese kann sich auf die oben genannten Ereignisse beziehen. Im Beispiel ist sie dem Tag DEAT DEAT zugeordnet, beschreibt also den Ort des Todes. Ortsangaben unter diesem Tag sollten die Verwaltungszugehörigkeit in aufsteigender Reihenfolge enthalten und mit einem Komma voneinander getrennt werden, beispielsweise: Stadt, Kreis, Bundesland, Land.[33]

Abb. 4: Auszug einer GEDCOM-Datei. [Goldberg 2022]

3. Entwicklung des (heuristischen) Algorithmus

Um historische Ortsangaben im deutschsprachigen Raum zu identifizieren und nachfolgend lokalisieren zu können, wird in diesem Abschnitt die Entwicklung eines Algorithmus beschrieben. In diesem wird die Heuristik zur kontextbasierten Auswahl eines Orts formalisiert. Zunächst wird die Metaheuristik erläutert, die dem zu entwickelnden Algorithmus zugrunde liegt. Die einzelnen Unteraspekte der Metaheuristik werden danach ausführlicher behandelt. Darunter fallen die Datenbereinigung, der Vergleich mit dem Mini-GOV, die Realisierung einer Ähnlichkeitserkennung, die Hinzuziehung von Kontextdaten und deren Alternativen sowie die Ermittlung historischer administrativer Zugehörigkeiten.

3.1 Metaheuristik

Auch der Algorithmus bildet die eingangs eingeführte inhaltliche Zweiteilung ab: Zum einen die Identifizierung eines Ortes, zum anderen die Bestimmung der historischen administrativen Zugehörigkeit zu einem manuell definierten Zeitpunkt (im Folgenden ›Bezugszeit‹ genannt). Die Identifizierung wird zudem in zwei Unterabschnitte getrennt: Im ersten Teil werden zunächst alle Ortsangaben des Kontextes gesammelt und bewertet.[34] Als Erstes werden dabei die Ortsangaben grundlegend bereinigt. Es wird je Ortsangabe geprüft, wie viele Übereinstimmungen mit den Bezeichnungen im Mini-GOV vorliegen. Folgend werden all die Orte, die genau *eine* exakte Übereinstimmung aufweisen, identifiziert und in die Kontextdaten mit aufgenommen (siehe Abbildung 5). Neben den Bezeichnungen und der GOV-ID werden zu den Kontextdaten auch die geographischen Koordinaten gespeichert.

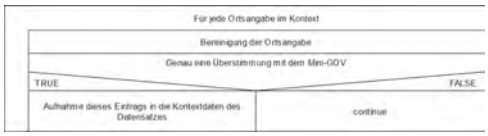

Abb. 5: Ermittlung und Bewertung von Kontextdaten, eigene Darstellung als Nassi-Shneiderman-Diagramm.[Goldberg 2022]

Mit Hilfe der nun ermittelten Kontextdaten können im zweiten Teil weitere Ortsangaben identifiziert werden (siehe Abbildung 6): Bei mehreren Übereinstimmungen mit dem Mini-GOV wird geprüft, ob für die zu überprüfende Ortsangabe Kontextdaten zur Verfügung stehen.[35] Falls ja, so werden für die verschiedenen Möglichkeiten jeweils die Koordinaten ermittelt und mit den Koordinaten der Kontextdaten verglichen. Der Ort, der nach diesem Vergleich die geringste Distanz aufweist, wird ausgewählt. Sind keine Kontextdaten vorhanden wird über eine Zuordnung auf Basis des Typs der möglichen Orte entschieden.

Liegt nach dem eingänglichen Vergleich mit dem Mini-GOV keine genaue Übereinstimmung vor, so findet eine Ähnlichkeitsüberprüfung zwischen der Ortsbezeichnung und den Einträgen im Mini-GOV statt. So können Fehler korrigiert werden, die durch die eingängliche Bereinigung nicht erkannt wurden. Besteht dennoch keine Ähnlichkeit mit einem Objekt aus dem Mini-GOV, bleibt der Ort unidentifiziert. Bei einem oder mehreren Treffern wird wie oben beschrieben verfahren.

---

[32] Neben dem Standard können einzelne programmspezifische Tags entwickelt werden. Diesen ist ein Unterstrich (›_‹) vorangestellt. So ist es prinzipiell möglich, weitere ortsbezogene Tags zu entwickeln und anzuwenden. Hervorzuheben ist hierbei insbesondere der Tag _LOC _LOC , über den ein eigener Datensatz zu der Ortsangabe erstellt wird. Darin können u. a. Tags wie _GOV _GOV genutzt werden, über welchen direkt die GOV-URI zugeordnet wird. Über die Tags LATI LATI und LONG LONG können zudem seit der GEDCOM-Version 5.5.1 direkt Koordinatenangaben gemacht werden (vgl. Gellatly 2015, S. 118). In der Validierung werden diese Tags nochmals aufgegriffen.

[33] Gellatly 2015, S. 117.

[34] Die Metaheuristik beschränkt sich bei der Hinzuziehung des Kontexts zur Identifizierung auf weitere Ortsangaben. Eine Erläuterung für diese Auswahl ist in Abschnitt 3.4 zu finden. Quellenspezifisch muss der Kontext jeweils definiert werden.

[35] Es kann sein, dass der Kontext so einen geringen Umfang hat, dass die vorhergehende Ermittlung der Kontextdaten kein Ergebnis brachte. Das ist hier konkret der Fall, wenn keine weiteren Ortsangaben vorliegen.

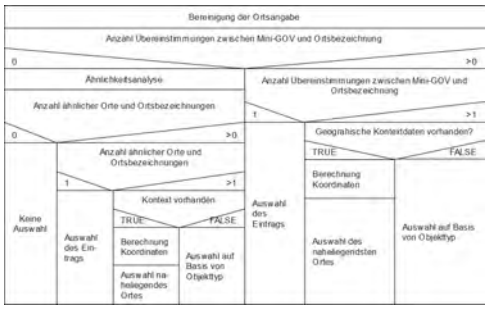

Abb. 6: Ortsidentifizierung, eigene Darstellung als Nassi-Shneiderman-Diagramm. [Goldberg 2022]

Der zweite Teil, die Clusterung, beginnt damit, dass für das zu untersuchende Objekt geprüft wird, ob eine Zuordnung zu einer historisch-administrativen Gliederungseinheit bereits früher erfolgt ist (siehe Abbildung 7). Das dient der Vermeidung doppelter Programmdurchläufe und somit der Laufzeitverkürzung. Wurde das Objekt noch keiner übergeordneten Einheit zugeordnet, so wird diese in der Baumstruktur der übergeordneten Objekte gesucht. Das erfolgt, indem in maximal zehn Iterationen jeweils ein übergeordnetes Objekt ausgewählt und neu untersucht wird. Hierzu wird jeweils geprüft, ob das Objekt Teil einer definierten Menge ist. Diese Menge stellt die historisch-administrative Zielgliederung (im Folgenden ›Provinz‹) zur Bezugszeit dar. Entspricht das Objekt einem Element der Zielmenge, so wird dem ursprünglichen Objekt diese Provinz zuordnet. Ist das nicht der Fall, so werden die Informationen aus dem GOV-Webservice für das entsprechende Objekt ausgelesen. Dieses Objekt kann nun wiederum für sich übergeordnete Objekte aufweisen, von denen eines für die nächste Iteration auszuwählen ist. Das geschieht, indem jedes übergeordnete Objekt dahingehend geprüft wird, ob die Bezugszeit in den Zeitraum der Zugehörigkeit des untergeordneten zum übergeordneten Objekt fällt. Ist das der Fall, wird das übergeordnete Objekt einer Liste A (höhere Priorität) hinzugefügt. Um beim obigen Beispiel zu bleiben, könnte etwa ›Mark Brandenburg‹ als übergeordnetes Objekt für ›Berlin‹ ausgewählt werden, wenn die Bezugszeit beispielsweise ›1301–1400‹ beträgt. Liegt die Bezugszeit hingegen nicht in der Zeitspanne der Zugehörigkeit zum übergeordneten Objekt, so wird das Objekt einer Liste B (niedrigere Priorität) zugewiesen.[36] Enthält die entsprechende Liste ein Element, so wird dieses Objekt ausgewählt und mit ihm die neue Iteration begonnen. Enthält die Liste jedoch mehrere Objekte, dann findet ein Vergleich dieser statt. Das ist überwiegend der Fall, wenn die Liste B Verwendung findet. Letztlich wird das Objekt ausgewählt, das zeitlich der Bezugszeit am nächsten ist.

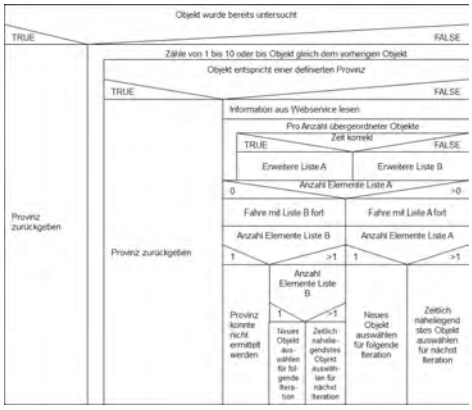

Abb. 7: Analyse der provinziellen Zuordnung, eigene Darstellung als Nassi-Shneiderman-Diagramm. [Goldberg 2022]

3.2 Datenbereinigung

Da es denkbar ist, dass Ortsbezeichnungen systematische Fehler enthalten, ist eine grundsätzliche Bereinigung der Daten angebracht.[37] Der Algorithmus sollte auf die jeweils auftretenden systematischen Fehler der Quelle spezifisch angepasst werden. Die Veränderung der Ortsbezeichnungen ist dabei nicht unkritisch, da mit dieser eine Interpretation der Daten einhergeht. Auf große Datenmengen angewendet kann es bei verallgemeinerten Korrekturregeln zu vermehrten Fehlinterpretationen kommen (FP). An dieser Stelle werden darum nur allgemeine Fehler abgedeckt (zum Beispiel die Entfernung von Sonderzeichen oder sonstiger unerwünschter Bestandteile). Spezifische Regeln können vom Anwender jeweils ergänzt werden.

[36] Liste A und B dienen der Priorisierung weiterer Schritte. Falls die Liste A kein Element enthält, ist kein übergeordnetes Objekt zeitlich exakt passend. Das kann u. a. daran liegen, dass das GOV nicht vollständig gepflegt ist oder aber ein Objekt erst nach der Bezugszeit entstanden ist und deswegen keine Zugehörigkeit zur Bezugszeit bestehen kann. Nur im Falle einer leeren Liste A wird mit der Liste B weiter verfahren.
[37] Je nach Ermittlung der Bezeichnungen sind dabei andere Fehler denkbar. So wird eine automatisierte Texterkennung andere Fehler machen als ein Mensch.

3.3 Suche

Grundlage der Suche eines passenden Ortes sind die Angaben über Namen von Orten im Mini-GOV – und der (teilweisen) Übereinstimmungen mit der gegebenen Ortsbezeichnung aus den Quellen. Das Mini-GOV liegt als CSV-Datei für jeden heute existierenden Staat vor. Aus diesem Grund müssen (neben dem Deutschland-Mini-GOV) weitere Länder, in denen es vormals deutschsprachige Gebiete gab, integriert werden. Die einzelnen Mini-GOVs werden zu einem gemeinsamen zusammengeführt. Dadurch ergibt sich eine Liste mit mehreren Hunderttausend Orten. Zur Reduzierung der Suchdauer des Algorithmus ist es hier wichtig, keine lineare Suche zu verwenden, sondern das Mini-GOV vorher zu sortieren.[38] Die Sortierung ermöglicht den Einsatz einer binären Suche.[39] Die Suche kann drei Ergebnisse hervorbringen:

- Kein Treffer: Es gibt keinen Wert im definierten Bereich des Mini-GOVs, der genau der bereinigten Ortsbezeichnung entspricht.
- Genau ein Treffer: Es gibt nur einen Wert im definierten Bereich des Mini-GOVs, der genau der bereinigten Ortsbezeichnung entspricht.
- Viele Treffer: Es gibt mehr als einen Wert im definierten Bereich des Mini-GOVs, der genau der bereinigten Ortsbezeichnung entspricht.

Bei genau einem Treffer wird eben dieser ausgewählt. In den anderen beiden Fällen ist eine weitere Analyse notwendig. Beim Ausbleiben eines Treffers werden ähnliche Zeichenketten gesucht, da u. a. Tippfehler in den Ortsbezeichnungen eine genaue Übereinstimmung verhindern. Im Fall mehrerer Treffer wird der Kontext der Ortsangabe zur weiteren Identifikation herangezogen (siehe Abschnitt 3.4).

Wird keine genaue Übereinstimmung zwischen Ortsbezeichnung und dem Namen eines Ortes im Mini-GOV gefunden, kann eine teilweise Übereinstimmung bei der Identifizierung helfen. Eine Berechnung zwischen der Ähnlichkeit aller Namen im GOV und der zuzuordnenden Ortsbezeichnung ist vor dem Hintergrund der Rechenkapazität nicht erstrebenswert. Aus diesem Grund werden lediglich die Orte verglichen, die zuvor schon eingegrenzt wurden.[40] Zum Vergleich von Strings existieren, wie in Abschnitt 2.2 erörtert, verschiedene Distanzmaße.[41] Die Verwendung unterschiedlicher Distanzmaße wird im Abschnitt 5. Validierung diskutiert.

3.4 Kontextdaten

Ansatzpunkte zur Ermittlung der Kontextdaten können sich aus der Quelle der zu lokalisierenden Ortsangabe ergeben. Als Kontext für eine Bezeichnung in einem Buch kann so unter Umständen das ganze Buch dienen. Wie in Abschnitt 2.1 gezeigt, gibt es verschiedene Informationen im Kontext von Ortsbezeichnungen (insbesondere andere Ortsbezeichnungen, Vor- und Nachnamen, Zeitangaben). Im Weiteren werden ausschließlich die anderen Ortsangaben als Kontextangaben genutzt. Der Ausschluss von Vor- und Familiennamen sowie Zeitangaben geschieht vor dem Hintergrund, dass zu den Namen einerseits keine Studie zur geografischen Verteilung ab dem 17. Jahrhundert existiert,[42] andererseits wird die geographische Zuordnung von Vornamen mit zunehmender Zeit unschärfer.[43] Eine fehlende Übersicht zur Veränderung der Ortsnamen im Zeitverlauf führt dazu, dass auch der temporale Aspekt nicht weiter herangezogen wird. Zudem betrifft die Veränderung der Bezeichnung innerhalb der Neuzeit nur einen kleinen Teil der Orte. Dagegen sind weitere Ortsangaben als Kontext besonders geeignet, vor allem wenn eine Dichte Nennung von Orten erfolgt. Bei einer Ortsangabe in einem Buch, in dem der nächste Ort erst 200 Seiten

[38] Beispielsweise bei einer GOV-Liste von 100.000 Orten und 10.000 zu lokalisierenden Urbanonymen wären bei einer einfachen Suche eine Milliarde einzelne Vergleiche durchzuführen.
[39] Diese wird so ausgeführt, dass zunächst der mittlere Eintrag der sortierten Instanz des Mini-GOVs identifiziert wird. Liegt die zu suchende Bezeichnung in alphabetischer Reihenfolge vor dem Ort, der in der Mitte steht, wird in einem nächsten Durchlauf die Mitte der vorderen alphabetischen Hälfte analysiert. Das wird so lange wiederholt, bis die nächste Mitte nur noch 10 Positionen von der vorherigen entfernt ist. Da es in einer alphabetischen Reihenfolge in einer Liste mehrere aufeinanderfolgende Treffer geben kann, ist eine vollständige Ausführung der binären Suche nicht angebracht. Diese würde nur ein Element zum Resultat haben. Aus diesem Grund werden die 30 Positionen vor und 30 Positionen nach der letzten ermittelten Mitte zusätzlich linear durchsucht. Das liegt darin begründet, dass kein einzelner Ortsname mehr als 30 Mal im GOV vorkommt.
[40] Das führt dazu, dass Fehler bei anfänglichen Buchstaben nicht erkannt werden. Die Vorteile durch die schnellere Suche scheinen jedoch für die praktische Durchführung wichtiger zu sein.
[41] Zandhuis et al. empfehlen zwar initial die Levenshtein-Distanz, bei mehreren Treffern sollte jedoch ein Vergleich der geographischen Entfernung vorgenommen werden. Zandhuis et al. 2015, S. 37.
[42] Allenfalls gibt es solche Betrachtungen für einzelne Regionen oder bestimmte Namen, beispielsweise in der Verteilung verschiedener Schreibweisen des Namens ›Mayer‹ (vgl. Hohensinner 2011).
[43] Die Namensgebung passiert heute in Deutschland weniger formalisiert als noch vor 100 Jahren, sondern bietet mehr individuelle Freiheiten (u. a. eine deutlich größere Variation der Vornamen).

später genannt wird, ist der Kontext hingegen zur Identifizierung voraussichtlich nicht geeignet. Sind in einer Quelle auf 20 Seiten aber 200 Ortsangaben, erscheint es wahrscheinlicher, dass diese in einer inhaltlichen Beziehung zueinander stehend genannt werden.[44] Vor der Anwendung dieses Algorithmus sollten eine Quelle und ihr Kontext dahingehend begutachtet werden.

Eine Eigenschaft dieser Vorgehensweise ist, dass Ortsangaben, sollen sie als Kontext zur Identifizierung anderer Ortsangaben genutzt werden, selbst zunächst einmal identifiziert werden müssen. Aus diesem Grund werden alle Ortsangaben zunächst auf Übereinstimmung mit dem Mini-GOV geprüft. Die Ortsangaben, bei denen eine eindeutige Zuordnung zu einem Eintrag im Mini-GOV vorgenommen werden kann, bilden die Kontextdaten. Hier wird deutlich, dass nur ein Teil der Ortsangaben in die Kontextdaten eingeht.

Um anhand der Kontextdaten nun eine Auswahl zwischen verschiedenen Orten zu treffen, ist die geographische Nähe das wesentliche Kriterium. Den Grundgedanken, dass die »geographical unit that has the smallest geographic distance to the place of origin« für die Zuordnung von Ortsangaben genutzt werden kann, hatten bereits Zandhuis et al.[45] Wobei der »place of origin« in dem Fall der Entstehungsort der Quelle der Ortsangabe ist, der (1.) selten vorliegen dürfte und (2.) auch erstmal identifiziert werden müsste. Der Ansatz hingegen, andere Ortsangaben desselben Kontextes zu nutzten, ist neu und basiert auf der Annahme, dass zusammen genannte Orte oftmals auch nah beieinanderliegen. Insbesondere bei genealogischen Quellen trifft diese Annahme zu. So werden oftmals Ehepartner aus einer engeren räumlichen Entfernung gewählt.[46] Oftmals wurden auch Ehepartner aus dem gleichen Ort gewählt.[47] Ein hoher Anteil der Bevölkerung war lange nur sehr eingeschränkt mobil. Diese Methodik ist also insbesondere geeignet für Urbanonyme, die zusammen mit Personen genannt werden (beispielsweise in prosopografischen Quellen). Weniger geeignet ist sie für Quellen, bei denen die Orte eindeutig in keinem räumlichen Zusammenhang stehen (z. B. eine alphabetische Liste aller Städte).

Zu jedem Ort in den Kontextdaten liegen durch das Mini-GOV Koordinaten vor. Das geographische Mittel[48] dieser Orte stellt dabei eine Position dar, die zu allen anderen Orten durchschnittlich die geringste Distanz hat. In dieser Vorgehensweise wird ein weiterer Nachteil ersichtlich: Die Quelle kann Orte enthalten, die geografisch sehr weit voneinander entfernt sind. Wenn diese sich nicht gleichflächig über den Raum verteilen (wovon auszugehen ist), können sich verschiedene Räume mit einer großen Dichte an Datenpunkten ausbilden.[49] Die geographische Mitte würde in keinem dieser Ballungsgebiete liegen. Hierbei schafft die Bildung dezentraler Cluster[50] Abhilfe. Die Validierung wird zeigen, welche praktischen Auswirkungen diese Art der Clusterung mit sich bringt.

Für den Fall, dass keine Kontextdaten vorhanden sind, jedoch trotzdem eine Auswahl zwischen verschiedenen Orten stattfinden soll, ist eine alternative Entscheidungsfindung zu definieren. So kann – unter Hinzuziehung weiterer Heuristiken – auf die Wahrscheinlichkeit für die Zugehörigkeit eines Urbanonyms zu einem bestimmten Ort geschlossen werden. Wenn z. B. gleichnamig eine Stadt und ein kleiner Weiler mit demselben Namen existieren, dann ist es wahrscheinlicher, dass die Stadt gemeint ist. Dahinter wiederum verbirgt sich die Annahme, dass mehr Menschen in einer Stadt als in einer kleinen Ansammlung von Häusern gelebt haben. Diese Unterscheidung wird durch die Objekttypen im Mini-GOV ermöglicht. Dazu wird folgende Reihenfolge definiert:

1. Kreisfreie Stadt
2. Stadt
3. Dorf
4. Pfarrdorf
5. Ort
6. Ortsteil
7. Ortschaft
8. Wohnplatz
9. Weiler

Sollten zwei Orte gleichen Objekttyps übrigbleiben, ist eine Identifizierung fehlgeschlagen.

[44] Auch diese Aussage ist nicht allgemeingültig: In einer alphabetisch sortierten Ortsliste ist das nicht der Fall.
[45] Zandhuis et al. 2015, S. 37.
[46] Vgl. Kocka et al. 1980.
[47] Vgl. Bähr et al. 1992.
[48] Es gibt keine allgemeingültige Norm des geographischen Mittels, da neben den Koordinaten beispielsweise die Topographie in die Gewichtung mit einfließen kann. An dieser Stelle wird das arithmetische Mittel jeweils von Längen- und Breitengrad genutzt.
[49] Eine alternative Idee ist es aus diesem Grund, nicht alle Orte in die Kontextdaten einfließen zu lassen, sondern vormals nochmal zu selektieren. Bei genealogischen Daten könnten z. B. nur die Orte der Personen, die eine nahe verwandtschaftliche Verknüpfung aufweisen, einbezogen werden. Die Verwandtschaftsbeziehungen stellen weitere Kontextdaten dar. Die wenigsten historischen Quellen verfügen jedoch speziell über diese Kontextdaten. Aus diesem Grund findet dieser Aspekt keine Anwendung.
[50] Diese Bildung von Cluster aus den Kontextdaten ist nicht mit der (späteren) Clusterung der identifizierten Orte entsprechen ihrer regionalen Zusammengehörigkeit identisch. Vielmehr dürfen diese nicht verwechselt werden. In beiden Fällen wird geclustert, weshalb der Begriff derselbe ist. Zur Unterscheidung ist ein genauer Blick auf den Zusammenhang zu werfen, in dem der Begriff verwendet wird.

3.5 Clusterung zur historischen administrativen Zugehörigkeit

Sind die Orte (im Folgenden ›Objekte‹) identifiziert, so stehen zu ihnen verschiedene Metadaten aus dem Mini-GOV zur Verfügung. Angaben zur historischen administrativen Zugehörigkeit sind dort allerdings nicht zu finden. Über den Webservice des GOV können diese Angaben jedoch ermittelt werden, da dort die historische Zugehörigkeit eingesehen werden kann. Die hierarchische Anordnung von unter- und übergeordneten Objekten ergibt eine Baumstruktur. Jedes Objekt hat Informationen über seine übergeordneten Objekte. In der Folge liegt die Baumstruktur nicht zu Beginn vor, sondern wird mit der Abfrage jedes übergeordneten Objektes implizit weiter aufgebaut. Durch die Verknüpfung der Zugehörigkeiten ergibt sich eine Baumstruktur (Beispiel in Abbildung 8). Um eine administrative Zugehörigkeit zu ermitteln, ist es Ziel, den richtigen Pfad in der Baumstruktur zu finden. Das geht nur, wenn zuvor mögliche Ziele (Cluster, Provinzen) definiert werden. Die Zielmenge ist dabei zeitabhängig: Je nach gewünschter Zeit kann die Clusterung anders ausgestaltet sein. Dazu muss die Zielmenge für verschiedene Bezugszeiten definiert werden. Fertig et al. strukturieren das deutschsprachige Gebiet in 39 Einheiten, die den Zeitraum von 1816 bis 1871 abdecken.[51] Für einen möglichen Bezug auf die heute existierende Gliederung werden die sechzehn Bundesländer Deutschlands genutzt.[52] Allerdings könnte auch eine Gliederung heutiger Daten in den Grenzen von 1850 – oder andersherum, die Einteilung historischer Daten in den heutigen Grenzen – interessant sein. Diese wird durch eine Variation der Bezugszeit ermöglicht.
Eine Zuordnung möglicher Provinzen mit GOV-Objekten ist folgend aufgeführt (siehe Tabelle 2). Weitere Regionen – oder eine detailliertere Skalierung der möglichen Cluster (u. a. Kreise, Regierungsbezirke) – können bei Bedarf implementiert werden.

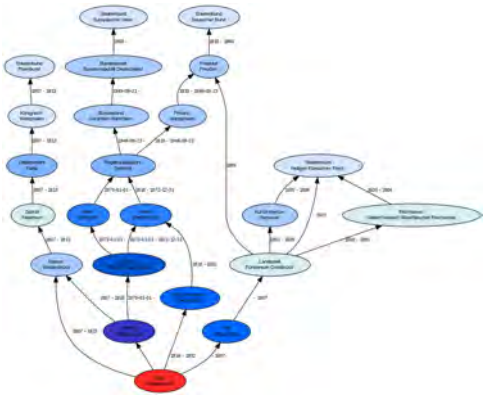

Abb. 8: GOV-Struktur der Stadt Wiedenbrück, Stand: 30. Januar 2021. [Goldberg 2022]

Provinz[53]	GOV-Identikator	Bemerkung
ab hier Bezugszeit <= 1871		
A 01 Provinz Holstein	object_190122	
A 02 Provinz Lauenburg	adm_131053	
A 03 Provinz Brandenburg (ohne Berlin)	object_1081716	
A 04 Provinz Hessen-Nassau	object_190330	
A 05 Provinz Hohenzollern	object_268785	
A 05 Provinz Hohenzollern	object_284443	Hohenzollern-Sigmaringen 1850 nach Hohenzollerschen Landen
A 06 Provinz Ostpreußen	adm_368500	
A 07 Provinz Pommern	adm_368480	
A 08 Provinz Posen	object_211667	
A 09 Provinz Sachsen	object_279654	
A 10 Provinz Schlesien	adm_368470	
A 11 Provinz Westfalen	object_190325	

Tab. 2: Zuordnung von Provinzen zum GOV. [Goldberg 2022]

[51] Vgl. Fertig et al. 2018.
[52] Beispiel: Bei einem Interesse an der Zuordnung Dresdens wäre im Jahr 2000 der Freistaat Sachsen korrekt (GOV-Objekt ›object_218149‹), 1850 aber beispielsweise das Königreich Sachsen (GOV-Objekt ›object_218149‹).
[53] Die mit A angeführten Provinzen sind Preußen zugehörig, dem mit dem B stellen weitere Territorien dar.

A 12 Provinz Westpreußen	object_213750	
A 13 Rheinprovinz	object_1047283	Provinz Jülich-Kleve-Berg bis 1822
A 13 Rheinprovinz	object_405464	Provinz Großherzogtum Niederrhein bis 1822
A 13 Rheinprovinz	object_190337	
A 14 Provinz Berlin	BERLINJO62PM	
B 01 Amt Bergedorf	object_257607	
B 02 Hansestadt Bremen	adm_369040	
B 03 Stadt Hamburg	adm_369020	
B 04 Stadt Lübeck	LUBECKJO53IU	
B 05 Stadt Frankfurt am Main	adm_136412	
B 06 Fürstentum Lippe-Detmold	object_217406	
B 07 Fürstentum Schaumburg-Lippe	object_217818	
B 08 Fürstentum Waldeck-Pyrmont	object_218152	
B 09 Großherzogtum Oldenburg	object_352387	
B 10 Großherzogtum Baden	object_217952	
B 11 Hessen	object_218147	
B 12 Großherzogtum Mecklenburg-Schwerin	object_217750	
B 13 Großherzogtum Mecklenburg-Strelitz (einschließlich des Fürstentums Ratzeburg)	object_217749	
B 14 Herzogtum Anhalt	object_190873	
B 15 Herzogtum Braunschweig	object_217954	
B 16 Herzogtum Nassau	object_218153	
B 17 Herzogtum Schleswig	object_190098	
B 18 Königreich Württemberg	object_190729	
B 19 Königreich Bayern	object_217953	
B 20 Königreich Hannover	object_190327	
B 21 Königreich Sachsen	object_218149	
B 22 Kurfürstentum Hessen	object_275299	
B 23 Landgrafschaft Hessen-Homburg	object_284442	
B 24 Thüringische Staaten	object_218143	Sachsen-Weimar-Eisenach
B 24 Thüringische Staaten	object_284441	Reuß Jüngere Linie
B 24 Thüringische Staaten	object_218134	Reuß Ältere Linie
B 24 Thüringische Staaten	object_218137	Sachsen-Altenburg
B 24 Thüringische Staaten	object_218138	Sachsen-Coburg-Gotha
B 24 Thüringische Staaten	object_265487	Sachsen Gotha
B 24 Thüringische Staaten	object_218142	Sachsen-Meiningen
B 24 Thüringische Staaten	object_218150	Schwarzburg-Rudolstadt
B 24 Thüringische Staaten	object_218151	Schwarzburg-Sondershausen
B 24 Thüringische Staaten	object_218141	Sachsen-Hildburghausen
ab hier Bezugszeit >= 1990		
Land Baden-Württemberg	adm_369080	
Freistaat Bayern	adm_369090	

Tab. 2: Zuordnung von Provinzen zum GOV. [Goldberg 2022]

Land Berlin[54]	BERLINJO62PM	
Land Brandenburg	adm_369120	
Freie Hansestadt Bremen	adm_369040	
Freie und Hansestadt Hamburg	object_1259992	
Land Hessen	adm_369060	
Land Mecklenburg-Vorpommern	adm_369130	
Land Niedersachsen	adm_369030	
Land Nordrhein-Westfalen	adm_369050	
Land Rheinland-Pfalz	adm_369070	
Saarland	adm_369100	
Freistaat Sachsen[55]	object_218149	
Land Sachsen-Anhalt	adm_369150	
Land Schleswig-Holstein	adm_369010	
Freistaat Thüringen	adm_369160	

Tab. 2: Zuordnung von Provinzen zum GOV. [Goldberg 2022]

Zur Suche in der Baumstruktur über- und untergeordneter Objekte wird im Folgenden ein informiertes Vorgehen genutzt. Dazu wird jeweils geprüft, ob das derzeitige Objekt Teil der Zielmenge ist. Ist es das nicht, werden die übergeordneten Objekte durchsucht. Hierbei wird überprüft, ob es ein Objekt gibt, welches in den Zeitraum der Bezugszeit fällt. Kann auf die Weise kein übergeordnetes Objekt identifiziert werden, wird dasjenige ausgewählt, dessen Grenzen der zeitlichen Zugehörigkeit am nächsten an der Bezugszeit liegen (z. B. die Bezugszeit 1820 und die zeitlichen Grenzen des nächsten Objektes von 1822-1838).[56] Sind ausschließlich übergeordnete Objekte ohne Zeitangaben vorhanden, weisen sie den gleichen Abstand zur Bezugszeit auf. In dem Fall wird die nächste Iteration mit dem erstgenannten Objekt der GOV-Abfrage durchgeführt. Solche Objekte, die einen kirchlichen oder juristischen Typ[57] aufweisen oder nur im nicht-deutschsprachigen Raum vorkommen, werden ausgeschlossen.[58] Wird ein übergeordnetes Objekt identifiziert, das Teil der Zielmenge ist, ist die Suche der historischen administrativen Gebietskörperschaft gelungen – der Ort wird der Provinz zugeordnet.[59]

Nach der Ermittlung der Zugehörigkeit zu einer historischen Provinz wird diese gespeichert. Bei einer großen Anzahl von zu identifizierenden Ortsbezeichnungen bietet diese Vorgehensweise den Vorteil, dass doppelt vorkommende Ortsangaben mit denselben Kontextdaten nicht doppelt bearbeitet werden.

Dieser Algorithmus findet die entsprechende Provinz nicht, wenn

1. durch die GOV-Abfrage keine übergeordneten Objekte ausgegeben werden (Beispielobjekt ›LIEHA2JO62RV‹) oder
2. kein Objekt der Zielmenge in den übergeordneten Objekten auftaucht,[60] was
3. auch daran liegen kann, dass der identifizierte Ort im Ausland (also außerhalb der Zielmenge) liegt.

[54] Berlin heute stellt das gleiche GOV-Objekt dar wie bei der historischen Klassifizierung. Insofern wir hier die Bezeichnung *A 14 Provinz Berlin* und nicht *Land Berlin* ausgegeben.
[55] Hier besteht die gleiche Situation wie bei Berlin (siehe vorherige Fußnote).
[56] Die Nähe der Grenzen zur Bezugszeit ist auch in dem seltenen Fall das Kriterium, in dem mehrere übergeordnete Objekte in der Bezugszeit liegen.
[57] Objekte können im Zeitverlauf verschiedenen Typen unterliegen, dieser Umstand wird nicht weiter betrachtet, weil sich die Objektkategorie nicht ändert. Ein Dorf kann beispielsweise zur Stadt werden, eine Kirche aber nicht zur Stadt.
[58] Beispielsweise bei dem Objekt ›NEUTE1JO44XB‹ gibt es zwei übergeordnete Objekte ohne zeitliche Angaben. Hier erfolgt jedoch eine konkrete Zuordnung zu dem Objekt, das keinen kirchlichen Typ aufweist.
[59] Eine Alternative dazu wäre eine klassische Tiefen- oder Breitensuche. Hierbei würden alle Zweige des Baumes analysiert, bis ein Objekt der Zielmenge gefunden wird. Das ist vor dem Hintergrund von Laufzeitaspekten ungünstig, da dazu sehr oft auf das nur online verfügbare GOV zugegriffen werden müsste und die Internet-Schnittstelle zu einer Verlängerung der Durchlaufzeit führte.
[60] Das ist derzeit bei allen Orten der ehemaligen Provinz Holstein der Fall, da diese nicht mit dem entsprechenden Objekt im GOV verknüpft sind. Allerdings kann dieses auch einzelne Orte anderer Provinzen betreffen, z. B. Wesel (›WESSELJO31HQ‹, Stand 18. Juni 2020). Dadurch, dass das GOV stetig erweitert wird, können die fehlenden Verknüpfungen in der Zukunft nachgeholt werden.

4. Anpassung auf GEDCOM-Dateien

Zur Vorbereitung der Validierung wurde der entwickelte Algorithmus in der Programmiersprache Python 3.7 umgesetzt. Der Programmcode ist im Online-Repositorium einsehbar. Er gliedert sich in verschiedene Funktionen, die zunächst einzeln erläutert und dann in ihrem Zusammenhang dargestellt werden. Die Kommentare im Programmcode beschreiben die konkrete Umsetzung des Algorithmus detailliert, sodass hier auf eine tief gehende Erläuterung verzichtet wird. Zum besseren Verständnis des Programms wird hier vielmehr der Aufbau beschrieben. Das Verständnis der Struktur hilft dabei, das Programm anzupassen. Grundlegend ist das Programm in die drei Bestandteile aufgeteilt, die auch schon zur Teilung des Algorithmus dienten (siehe Abbildung 9). Hierzu werden in jedem Schritt eigene CSV-Dateien mit den Zwischenergebnissen erstellt und ausgegeben.

Abb. 9: Aufbau des Programms. [Goldberg 2022]

Vor der Main-Funktion wird dieser Prozess einmal je GEDCOM-Datei angestoßen (siehe Abbildung 10, die Pfeile zeigen an, in welcher Systematik die Funktionen aufgerufen werden). Bevor das geschehen kann, werden die zu untersuchenden GEDCOM-Datei vorbereitend so verändert, dass die Dateinamen eine Zahl gefolgt von der Dateinamen-Endung ›.ged‹ beinhaltet (d. h. ›1.ged‹, ›2.ged‹ etc.). Zahlen dürfen nicht doppelt vorkommen, jedoch ausgelassen werden. Eine durchlaufende Nummerierung ist empfohlen.

Über die Funktion `importMiniGOV()` `importMiniGOV()` findet ein Import des Mini-GOV statt. Hier runter sind verschiedene Textdateien zu verstehen, die zuvor je (heutigem) Staat vom CompGen bereitgestellt werden.[61] Es werden die Mini-GOV-Dateien zu Deutschland, Polen, sterreich, der Schweiz, Tschechien, Dänemark, Frankreich und den Niederlanden integriert. Deutschsprachige Ortsnamen sind als *aktueller Name* oder *letzter deutscher Name* gekennzeichnet. Falls es Letzteren gibt, so wird ein zusätzlicher Eintrag kreiert, indem der alte Name den neuen überschreibt.[62] Ansonsten werden die Mini-GOVs aneinandergesetzt und alphabetisch sortiert. Hintergrund der Sortierung ist die Ermöglichung einer binären Suche darin.

Da viele GEDCOM-Dateien einzeln und voneinander unabhängig zu bearbeiten sind, ist eine Parallelisierung des Programmcodes sinnvoll. Die Funktion `parallel()` `parallel()` wird dazu parallel aufgerufen und bearbeiten. Wesentlich darin ist der nacheinander folgende Aufruf von `mainMetadataInspector()` `mainMetadataInspector()`, `mainPlaceFinder()` `mainPlaceFinder()`und `mainProvinceFiner()` `mainProvinceFiner()`. Die Ergebnisse aus diesen jeweiligen Funktionsaufrufen werden mit der Funktion `appendFile()` `appendFile()` den CSV-Dateien hinzugefügt. Diese Dateien sind zuvor über die Funktion `createFile()` `createFile()` erstellt worden. Dazu dient unterstützend die Funktion `loadData()` `loadData()`. Ebenso wurden zuvor die Daten der jeweiligen GEDCOM-Datei über `loadGedcomFile()` `loadGedcomFile()` geladen.

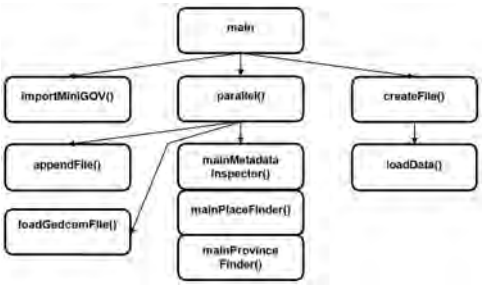

Abb. 10: Funktionsweise der Main. [Goldberg 2022]

Für die Inspektion des Kontextes wird eine Liste von Metadaten zu den einzelnen GEDCOM-Dateien geschaffen (siehe Abbildung 11). Dazu werden in der Funktion `prePlaceFinder()` `prePlaceFinder()` die Ortsangaben über den Tag `PLAC` `PLAC` erkannt und über die Funktion `minigovSearch()` `minigovSearch()` geprüft, ob sie nur eine Übereinstimmung im Mini-GOV

[61] Vgl. Verein für Computergenealogie e. V. (Hg.) 2021b.
[62] Die Eintragungen der anderen europäischen Staaten, die keinen letzten deutschen Namen aufweisen, werden bewusst nicht gelöscht, da das Programm unnötig verschlechtert würde und so ein Ansatz gelegt ist, um auch auf internationale Ortsangaben adaptiert zu werden.

aufweist. Hierzu wird über eine binäre Suche nach der Anzahl an Übereinstimmungen gesucht. Die jeweilige Ortsangabe wird dadurch bereinigt, indem nur der Teil nach einem möglichen ersten Komma entfernt wird. Wenn nur ein Treffer erzielt wurde, so wird der Ort in eine Liste eindeutiger Orte mit aufgenommen – die selektierten Kontextdaten.

Nachfolgend wird in der Funktion `qualityChecker()` `qualityChecker()` zunächst geprüft, ob die untersuchte Datei schon einmal vollständig verarbeitet wurde. In dem Fall wird für jede Ortsangabe zunächst nochmal geprüft, ob sie in der Liste der eindeutigen Werte vorhanden ist. Nicht zu allen eindeutigen Bezeichnungen gibt es jedoch Koordinatenangaben. Einträge, bei denen diese fehlen, finden keine weitere Beachtung. Mit den restlichen wird ein Clusterungsverfahren durchgeführt. Das gröbste mögliche Cluster stellt dabei den Mittelpunkt aus allen selektierten Kontextdaten dar. Sinnvolle Parameter für die Feinheit der Clusterung werden in der Validierung ermittelt. Die Koordinaten der Cluster werden zurückgegeben und ergänzen die Metadaten der Datei. `qualityChecker()` `qualityChecker()` greift auf verschiedene Funktionen zu: Mithilfe der Funktion `gedcomRowParser()` `gedcomRowParser()` werden die Bestandteile einer GEDCOM-Zeile separiert; `stringDoublingCounter()` `stringDoublingCounter()` dient der Zählung von Clustern.

Im Ergebnis bestehen die Metadaten aus der Bezeichnung der Datei, den Durchschnittskoordinaten der selektierten Kontextdaten, der Anzahl eindeutiger Ortsbezeichnungen, der Anzahl daraus gebildeter Cluster sowie den Mittelpunkten jedes einzelnen Clusters. Die durchschnittlichen Koordinaten ergeben sich aus dem arithmetischen Mittel der Längen- und Breitengrade aller so gefundenen Ortsangaben.

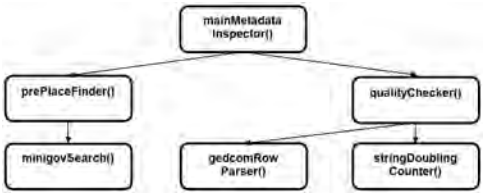

Abb. 11: Funktionsweise der Kontextdatenanalyse. [Goldberg 2022]

In der Funktion `mainPlaceFinder()` `mainPlaceFinder()` werden die Daten durch die Funktion `dataCleaner()` `dataCleaner()` zunächst einer weitreichenden Bereinigung unterzogen (siehe Abbildung 12). In der Funktion `stringFunc1()` `stringFunc1()` wird dazu unter anderem geprüft, ob eine bestimmte Zeichenkette am Anfang steht; diese wird sodann gelöscht, andernfalls wird alles vor dem Teilstring beibehalten. In `stringFunc2()` `stringFunc2()` fehlt diese alternative Bedingung. Wie im vorherigen Abschnitt zerteilt die Funktion `gedcomRowParser()` `gedcomRowParser()` die einzelnen GEDCOM-Zeilen in ihre Bestandteile.

Die konkrete Auswahl eines Orts zu einer gegeben Ortsbezeichnung auf Basis der Kontextdaten findet in der Funktion `find()` `find()` statt, die durch die Funktion `placefinder()` `placefinder()` vorbereitet wird. Hier wird im näheren der in Abbildung 6 gegebene Prozess umgesetzt. Unter anderem findet die kontextsensitive Zuordnung von Ortsbezeichnungen und Orten hier statt. Die Suche nach Clustern im Umkreis der Koordinate ist in die Funktion `areaSearch()` `areaSearch()` ausgegliedert. Hier wird auch die typenbasierte Zuordnung realisiert, falls die Umkreissuche fehlschlägt.

Als Ergebnis steht eine Tabelle zur Verfügung, die für jede (unbereinigte) Ortsbezeichnung in jeder Quelle eine Zeile mit der GOV-URI, ihren ermittelten Koordinaten, der Information, über welche Art und Weise die Funktion `find()` `find()` die Zuordnung getroffen hat, der bereinigten und unbereinigten Ortsangabe sowie den Dateinamen enthält.

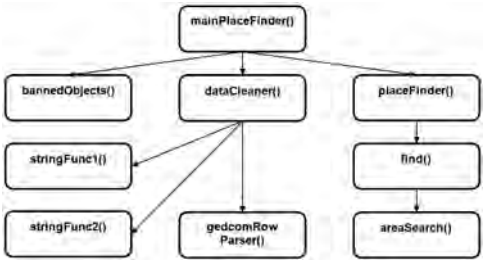

Abb. 12: Funktionsweise der Ortssuche. [Goldberg 2022]

Zuletzt wird die Provinzsuche durch die Funktion `mainProvincefinder()` `mainProvincefinder()` realisiert (siehe Abbildung 13). Zu jedem zuvor identifizierten Objekt wird die Funktion `provincefinder()` `provincefinder()` initiiert. Hierin findet die Suche in der übergeordneten Baumstruktur des GOVs statt. Die dazu notwendigen Informationen werden

über die Funktion `callWebservice()` `callWebservice()` aus dem Internet abgerufen; es ist also eine Internetverbindung vom ausführenden Gerät vonnöten. Es wird angenommen, dass kein Baum mehr als zehn Ebenen besitzt. Deswegen wird in maximal zehn Iterationen immer eines der übergeordneten Objekte ausgewählt und untersucht. Zunächst wird dazu geprüft, ob eines der Objekte bereits Teil der Zielmenge ist.[63] Wenn das nicht der Fall ist, wird die Zeitspanne der Zugehörigkeit zur Auswahl herangezogen. Zur Berechnung von Beginn- oder Endjahren aus Zeitspannen der Zugehörigkeit zu übergeordneten Objekten im GOV werden die Funktionen `beginCalculator()` `beginCalculator()` bzw. `endCalculator()` `endCalculator()` genutzt. In diesem wird aus der julianischen Angabe der Zeitspanne eine gregorianische Jahresangabe ermittelt. Da die Zeitangaben im GOV oftmals unvollständig sind, existiert ein System, dass mögliche Objekte priorisiert und so die Wahrscheinlichkeit vergrößert, das richtige Objekt auszuwählen. Dabei existieren auszuschließende Objekte (hier: Objekte kirchlichen Typus, ausschließlich Verwaltungsgliederungen anderer Staaten oder gerichtliche Institutionen), die nicht ausgewählt werden sollen. Hieraus folgt das Endergebnis: Eine Tabelle, die die ursprüngliche Bezeichnung, den Namen der Quelldatei, die GOV-URI sowie die Bezeichnung der zugeordneten Provinz enthält.

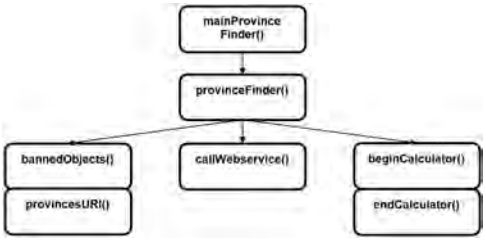

Abb. 13: Funktionsweise der Provinzsuche. [Goldberg 2022]

5. Validierung und Diskussion

Zur Validierung des Algorithmus werden GEDCOM-Dateien aus GEDBAS genutzt. Hier können GEDCOM-Dateien von Nutzer*innen ohne grundsätzliche Beschränkung hochgeladen werden. Dabei können die Nutzer*innen beim Hochladen zulassen, dass ihre Datei anderen zum Download frei zur Verfügung steht. Diese downloadbaren Dateien stellen die Datenbasis für die Validierung dar. Dazu wird der bei Goldberg und Moeller beschriebene Scraper genutzt.[64] Eine Ausführung des Scrapers am 15. April 2020 erbrachte 1.899 GEDCOM-Dateien. Diese enthalten 3.371.825 durch den Tag `PLAC` `PLAC`[65] definierte Ortsangaben. Eine Separierung der Verwaltungszugehörigkeit durch ein Komma wird in den GEDCOM-Dateien in den überwiegenden Fällen nicht eingehalten, sodass diese Regelung schwerlich zu Identifizierung genutzt werden kann – und da wo sie eingehalten wird, werden unterschiedliche Verfahren genutzt.[66]

Etwa 12 Prozent dieser Ortsbezeichnungen sind einem Objekt im Mini-GOV direkt zuzuordnen. Sie stellen demnach die Kontextdaten dar. Für weitere 34 Prozent gibt es mehrere, für 54 Prozent keinen Treffer. Diese 12 Prozent bilden die selektierten Kontextdaten, mit denen eine Clusterung durchgeführt werden kann. Für diese Anforderungen ist eine Clusterung gemäß DBSCAN-Verfahren geeignet (Density-Based Spatial Clustering of Applications with Noise). Der von Ester et al. beschriebene Algorithmus besagt, dass jeder Datenpunkt für sich betrachtet werden soll; ist er noch nicht klassifiziert wird die Bildung eines neuen Clusters oder die Zuordnung zu einem bestehenden angestoßen.[67] In Anlehnung daran ist eine Clusterung auch für die Kontextdaten realisiert (siehe Funktion `qualityChecker()` `qualityChecker()`). Um die Parameter der Clusterung anzupassen ist eine Betrachtung der Streuung der Cluster in einzelnen GEDCOM-Dateien von Interesse. Zum einen kann die Mindestanzahl von Orten variiert werden, die nötig sind, um ein Cluster zu bilden. Zum anderen ist die Mindestdistanz entscheidend, die ein Ort von einem Cluster entfernt sein muss, um diesem nicht zugeschlagen zu werden. Im Folgenden wird dieses für die GEDCOM-Dateien untersucht.[68] Bei einem Vergleich der Mittelpunkte von Clustern einer Quelle bei Variation der Parameter (Kilometer zur Zusammenführung zweier Orte, Mindestanzahl von Orten für ein Cluster, siehe Abbildung 14) zeigt

[63] Die Zuordnung von Provinzen zu GOV-Objekten ist in Tabelle 6 definiert. Diese GOV-URIs bilden die Zielmenge für die Clusterung der identifizierten Orte.
[64] Vgl. Goldberg / Moeller 2021.
[65] Neben dem Tag `PLAC` `PLAC` gibt es auch die Tags `FORM` `FORM` (wenn er sich auf `PLAC` `PLAC` bezieht) oder `_LOC` `_LOC`, ja sogar die direkte Möglichkeit der Angabe von Koordinaten über `LATI` `LATI` und `LONG` `LONG`. Ebenso existiert der Tag `_GOV` `_GOV`, mit dem zu einem Ort direkt die GOV-URI zugeordneten werden kann. Diese finden allerdings in GEDBAS eher weniger Verwendung (`FORM` `FORM` in Bezug auf `PLAC` `PLAC`: 2.085 Verwendungen über alle Dateien (diese allerdings in nur 8 verschiedenen Dateien), `_LOC` `_LOC`: 101.603 (50), `LATI` `LATI`: 54.719 (88), `LONG` `LONG`: 54.718 (88), `_GOV` `_GOV`: 5.043 (26)). Das Ergebnis bestätigt die Auffassung von Gellatly, dass diese Art der Geokodierung mehrheitlich nicht genutzt wird, vgl. Gellatly 2015, S. 118. Aufgrund dieser sehr geringen Verwendung wird sich im Weiteren auf den Tag `PLAC` `PLAC` bezogen. Es ist nachdrücklich erstrebenswert, dass Genealogen diese Felder künftig direkt pflegen.
[66] In 1.629.274 `PLAC` `PLAC`-Angaben (48 Prozent) kommen insgesamt 3.933.251 Kommata vor – bei einer konsistenten Verwendung wären es etwa zehn Millionen. Allerdings ist zu erkennen, dass die Kommasetzung in einigen großen (meist englischsprachigen) Dateien stark konzentriert ist: 80 Prozent der Kommata fallen in 2,6 Prozent der Dateien an. Allein 13 Dateien (von 2.899) enthalten die Hälfte aller `PLAC` `PLAC`-Kommata. Das zeigt deutlich, dass diese Ordnungssystematik bei der Identifizierung von GEDBAS-Urbanonymen keine breite Hilfestellung sein kann.
[67] Vgl. Ester et al. 1996, S. 229.
[68] Bei der Validierung durch andere Quellen würden gegebenenfalls andere Ergebnisse resultieren.

sich, dass eine Distanz von 10 km (links) dazu führt, dass es sehr viele Cluster gibt, die nah beieinander liegen. Dieses ist ein unerwünschter Effekt, da diese Cluster aus der Perspektive des gesamten deutschen Sprachraumes zu wenig voneinander entfernt sind. Es deutet sich ein weiteres lokales Maximum bei der Distanz von etwa 450 km an. Auch bei 25 km zeigt sich dieses Bild (Mitte), allerdings in sehr abgeschwächter Form. Erst bei 50 km ist der starke Anstieg nach Erreichen des Mindestabstandes abgeflacht (rechts). Der Erwartungswert der Verteilung liegt etwa bei 450 km. Gleiches gilt bei einer Variation der Mindestanzahl an Koordinatendatenpunkten je Cluster zu erkennen. Hierdurch wird jedoch die absolute Zahl der Cluster stark geschrumpft.

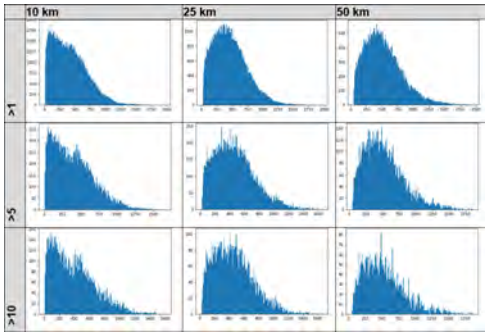

Abb. 14: Anzahl an Clustern nach durchschnittlichem Abstand in km bei Variation des Mindestabstandes zwischen Clustern (10 km, 25 km, 50 km) und Mindestgröße von Clustern (2, 6, und 11), jeweils 200 Bins. [Goldberg 2022]

Bei einer Betrachtung der Anzahl von Clustern pro Datei für jede dieser neun Optionen ergibt sich, dass die unterschiedlichen Optionen vor allem bei Dateien mit vielen Clustern eine Auswirkung aufzeigen. Liegen eine geringere Mindestgröße und ein geringer Minimalabstand zwischen zwei Punkten von Clustern vor, erhöht sich die Anzahl der gesamten Cluster stark (siehe Abbildung 15). Das bedeutet, dass bei einer niedrigen Mindestdistanz sehr viele nah beieinanderliegende Cluster existieren. Aus diesem Grund wurde die Mindestdistanz von 50 km gewählt. Als Mindestgröße eines Clusters wird >5 gewählt, um einerseits die starke Erhöhung der Clusteranzahl bei einem weiteren Absenken der Mindestgröße zu vermeiden, auf der anderen Seite aber auch in wenig umfangreichen Dateien eine Clusterbildung zu ermöglichen.

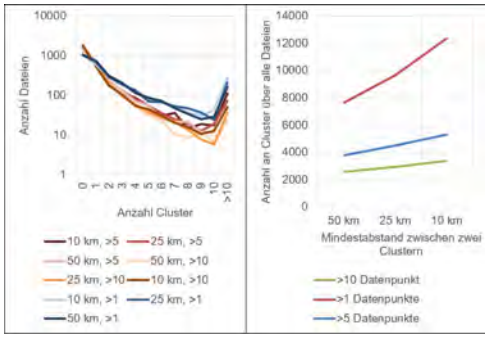

Abb. 15: Zusammenhang zwischen Mindestdistanz, Mindestclustergröße und Clusteranzahl, links mit logarithmischer Skalierung. [Goldberg 2022]

Definiert wird also, dass ein Cluster mindestens aus sechs Orten bestehen muss. Ferner muss jeder Ort eines Clusters 50 Kilometer von dem Ort eines anderen Clusters entfernt liegen; ansonsten würden sie in einem gemeinsamen Cluster zusammengeführt. Es zeigt sich, dass dadurch 66 Prozent der GEDCOM-Dateien, die überhaupt ein Cluster aufweisen,[69] nur ein einziges Cluster aufweisen, das den oben definierten Eigenschaften (50 km, mind. 6 zusammenhängende Orte) entspricht. 19 Prozent jedoch weisen zwei Cluster auf, 15 Prozent sogar mehr als zwei. Die Entfernung der Optionen einer unidentifizierten Ortsangabe werden für alle Clustermittelpunkte berechnet. Der Ort mit der geringsten Distanz zu einem der Clusterpunkte wird für die Identifizierung gewählt.

Ein weiterer Bestandteil der Validierung bezieht sich auf die Ähnlichkeitsanalyse. Diese ist zwar nicht primär zum kontextbasierten Aspekt des Algorithmus zu zählen, wird aufgrund der Relevanz für das Gesamtergebnis aber trotzdem diskutiert. Die Ortsangaben, zu denen keine Übereinstimmung gefunden werden konnten, werden einer Ähnlichkeitsanalyse unterzogen. Dazu werden zwei Varianten verglichen. Zum einen ist das die Levenshtein-Distanz (bereits in Abschnitt 2.2 erläutert). Zum anderen wird ein Maß gesetzt, dass sich hierarchisch auf verschiedene Bestandteile stützt: Zunächst werden vertauschte Zeichen erkannt, indem die 61 Orte[70] einer Anagrammdetektion unterzogen werden.[71] Falls kein Anagramm erkannt wird, werden

[69] 52 Prozent GEDCOM-Dateien weisen kein Cluster auf. Das liegt darin begründet, dass kein Cluster von mindestens fünf Ortsangaben gebildet werden konnte.
[70] Siehe hierzu Anm. 41.

zunächst aufeinanderfolgende doppelte Buchstaben in beiden Zeichenketten entfernt. Sind diese dann gleich, besteht eine Zuordnung. Ist das nicht der Fall, so wird mit dem verbleibenden Zeichen ein Vergleich nach der Kölner Phonetik angestellt. Bei einem gleichen Ergebnis findet eine Auswahl statt.

Bei einem initialen Grenzwert von 0,25 (Verhältnis von Anzahl der Änderungen zur Länge der Zeichenkette[72]) ergeben sich sowohl bei der Levenshtein-Distanz als auch bei dem alternativen Maß viele Falschpositive (75 bzw. 74 Prozent). Ein Großteil dieser ist zum einen durch Ortsangaben zu Orten außerhalb des Betrachtungsgebietes bedingt, die dennoch deutschen Ortsnamen ähneln. Zum anderen sind dort überproportional viele Bezeichnungen vorhanden, die Orte in den ehemaligen Ostgebieten beschreiben und bei denen die zeitliche Konsistenz der Ortsbezeichnung (u. a. aufgrund von Umbenennungen) gering ist. Auch liegt das darin begründet, dass Ortsbezeichnungen teilweise sehr ähnlich sind.[73] Vermeintlich korrekte TP-Ergebnisse (z. B. ›Stuttgardt‹ zugeordnet zu ›Stuttgart‹) sind in der Minderheit. Aus diesem Grund wird die Grenze auf 0,17 (ab sechs Buchstaben eine Veränderung, ab 12 Buchstaben zwei Veränderungen etc.) verringert. In der Folge kommt die Ähnlichkeitsanalyse über die Levenshtein-Distanz etwa bei jedem sechsten Fall zu einem Ergebnis, das alternative Maß in jedem fünften Fall (siehe Tabelle 3). Bei 9,34 bzw. 11,90 Prozent ergeben sich mehrere Übereinstimmungen, von denen eine ausgewählt wird. Beide Ähnlichkeitsanalysen produzieren mehr falsche als richtige Ergebnisse. Dieses zeigt, dass die Ähnlichkeitsanalyse insbesondere bei einer schlechten Datenqualität und vor dem Hintergrund einer großen Ähnlichkeit von Ortsbezeichnungen eine Herausforderung ist, die hier nicht zur vollkommenen Zufriedenheit gelöst werden kann. Künftige Forschungen sollten sich diesem Aspekt verstärkt annehmen. Aber auch wenn der Wert bezogen auf die GEDCOM-Dateien nicht zufriedenstellend ist, so sollte dieses Element im Algorithmus dennoch beibehalten werden. Es ist zu erwarten, dass bei anderen Quellen mit einer besseren Datenqualität auch eine bessere Erkennung einhergeht.[74] Da der alternative Vergleich über Anagramme und die Kölner Phonetik bessere Werte (mehr TP-Ereignisse[75]) aufweist, wird diese Alternative im Weiteren ausgewählt.

Ähnlichkeitsmaß	Erkennung = 1	FP	TP	Erkennung > 1[76]	FP	TP
Levenshtein-Distanz	6,23 %	68	32	9,34 %	71	29
Anagramm und Kölner Phonetik	7,22 %	65	35	11,90 %	67	33

Tab. 3: Vergleich verschiedener Ähnlichkeitsmaße, alle Angaben in Prozent, Grenze: 0,17. [Goldberg 2022]

Eine Ausführung des Algorithmus unter den oben festgelegten Variationen ergibt das in Tabelle 4 dargestellte Ergebnis. 49,72 Prozent der Urbanonyme wird ein GOV-URI zugeordnet, 50,28 Prozent hingegen nicht. Der prozentuale Anteil der jeweils möglichen Ereignisse zeigt, dass fast jedes dritte Urbanonym dem weiteren Verfahren entzogen wird, weil es unerlaubte Bestandteile enthält. Hierzu dienen insbesondere Hinweise auf eine Angabe außerhalb des Betrachtungsgebiets (z. B. Kürzel von US-Bundesstaaten). Die kontextbasierte Auswahl auf Basis der geographischen Nähe bekommt daher eine hohe Relevanz. Auffällig ist auch der hohe Anteil an Ortsangaben, zu denen kein Pendant (also auch keine Ähnlichkeit) zu den Orten des Mini-GOVs vorhanden ist. Dem zugrunde liegen viele Berufsangaben und Orte außerhalb des Betrachtungsgebiets, aber auch vermeintlich korrekte Ortsangaben, die in einem unüblichen Format vorliegen (z. B. als Wikipedia-Verlinkung) oder sehr falsch geschrieben sind. Ansonsten finden sich auch hier überproportional viele Angaben, die auf ehemalige deutsche Ostgebiete hinweisen und nicht in der Ausführlichkeit im Mini-GOV gepflegt sind. Wenig relevant sind dagegen die Fälle, die die Typen der Objekte einbeziehen. Von diesen ist einzig allein die Option, bei der nur ein Objekt einen passenden Typ besitzt, mit 7,44 Prozent bedeutend.

Auswahlkriterium	Anteil in Prozent
Unerlaubte Bestandteile	31,51

Tab. 4: Endzustände des Algorithmus am Beispiel Ortsangaben in GEDCOM-Dateien, Grundgesamtheit von 262.741 Einträgen.[77] [Goldberg 2022]

[71] Der Nachteil dieses Vorgehens liegt auch hier darin, dass vertauschte Zeichen am Anfang der Zeichenkette nicht erkannt werden, weil sie in einer alphabetischen Sortierung weit entfernt sind. Zudem erkennt eine Ähnlichkeitsanalyse durch ein Anagramm nur vertauschte Zeichen, nicht aber ausgelassene oder zusätzlich vorhandene Buchstaben.
[72] Bei der Kölner Phonetik wird dazu die Levenshtein-Distanz der ursprünglichen Zeichenketten herangezogen.
[73] Bei einer zulässigen Levenshtein-Distanz von 2 sind beispielsweise Orte wie Nehmten und Nehren, Neiden und Nehren, Bövingen und Böttingen, Muron und Murr, Murow und Murr, Balden und Belsen, Weißstein und Beilstein oder Hattingen und Böttingen als gleich anzusehen. Selbst bei einer Verringerung der Toleranz auf eine Distanz von 1 würden viele FP-Ergebnisse produziert, beispielsweise bei Nehden und Nehren, Bötzingen und Böttingen oder Oderthal und Odenthal.
[74] Eine Idee für die künftige Erweiterung des Algorithmus besteht darin, dass unklare Ortsangaben zunächst innerhalb des Kontextes verglichen werden. Insbesondere in GEDCOM-Dateien besteht eine hohe Wahrscheinlichkeit, dass Orte doppelt genannt werden; Schreibfehler könnten hierüber gegebenenfalls effektiv erkannt werden.
[75] Zur Ermittlung wurden 100 zufällige Ergebnisse ausgewählt und manuell eingeschätzt, ob es sich um ein richtiges oder falsches Ergebnis handelt.
[76] Die Angabe zu den Falschpositiven und Falschnegativen bezieht sich zudem auf die nachfolgende Auswahllogik eines der erkannten Orte (anhand geographischer Nähe, Typ etc.); nur dieses wird untersucht.
[77] Die erhebliche Verkleinerung der Anzahl zu untersuchender Ortsangaben (von 3.057.810 auf 262.741) ist auf die Löschung doppelter Ortsangaben in einer Datei zurückzuführen.

Unerlaubter Angabe	0,07
Einziger passender Treffer in Ähnlichkeitsanalyse	2,05
Einziger mit Koordinaten nach Ähnlichkeitsanalyse	0,58
Geographische Nähe nach Ähnlichkeitsanalyse	2,42
Einziger gültiger Typ nach Ähnlichkeitsanalyse	0,02
Keiner mit gültigem Typ nach Ähnlichkeitsanalyse	0,00
Ein passender Typ nach Ähnlichkeitsanalyse	0,05
Zu viele passende Typen nach Ähnlichkeitsanalyse	0,00
Ein einziger passender Treffer	11,25
Ein einziger Treffer mit Koordinaten	6,53
Geographische Nähe	18,87
Einziger gültigen Typ	0,41
Keiner mit gültigem Typ	0,05
Ein passender Typ	7,44
Zu viele passende Typen	0,00
Kein Auswahlkriterium entscheidend	0,48
Kein Pendant im Mini-GOV gefunden	18,17

Tab. 4: Endzustände des Algorithmus am Beispiel Ortsangaben in GEDCOM-Dateien, Grundgesamtheit von 262.741 Einträgen.[77] [Goldberg 2022]

Eine Stichprobe von 100 Werten erbrachte 36 TP-Werte[78], 42 TN-Werte, 11 FP-Werte[79] und 11 FN-Werte.[80] Der hohe Anteil an Falschnegativen ist dadurch zu erklären, dass das Mini-GOV nicht alle Orte oder Varianten enthält (insbesondere in ehemaligen deutschsprachigen Siedlungsgebieten). Die Gründe für eine wahrnegative Lokalisierung hingegen sind überwiegend dadurch bedingt, dass Ortsangaben außerhalb des Betrachtungsgebiets als solche erkannt werden. Insgesamt sind also 78 Prozent der Ortsangaben korrekt und 22 Prozent nicht korrekt behandelt worden.

Der zweite Teil der Validierung bezieht sich auf die Clusterung der GOV-URIs zu den Provinzen. Bei insgesamt 68.011 GOV-URIs wurde versucht, mit dem Algorithmus eine Provinz zum Jahr 1820 zuzuordnen. Bei 46.877 GOV-URIs (69 Prozent) war die Zuordnung erfolgreich. Der hohe Anteil an nicht zugeordneten GOV-URIs ist jedoch diskussionswürdig. Aus diesem Grund wurden 100 zufällige nicht zuordenbare GOV-URIs manuell überprüft.[81] Die Fehler sind in Tabelle 5 dargestellt.

Grund der fehlgeschlagenen Klassifizierung	Anzahl
Ort liegt außerhalb der Zielprovinzen (In heutigen Staaten: Niederlande 29x, Frankreich 12x, Österreich 9x, Tschechien 5x, Schweiz 3x, Polen 3x, Dänemark 1x)	62

Tab. 5: Fehler in der Zuordnung der Provinzen bei der Bezugszeit 1820. [Goldberg 2022]

[77] Die erhebliche Verkleinerung der Anzahl zu untersuchender Ortsangaben (von 3.057.810 auf 262.741) ist auf die Löschung doppelter Ortsangaben in einer Datei zurückzuführen.
[78] Erkannte Orte aus den nicht-deutschen Mini-GOVs fallen in diese Kategorie.
[79] Nicht erkannte Orte aus den nicht-deutschen Mini-GOVs fallen in diese Kategorie.
[80] Die für die Einordnung in die vier Kategorien notwendige manuelle Klassifizierung basierte dabei auf der Hinzunahme von Informationen, die bei der Bereinigung gelöscht wurden sowie dem Aufrufen der GEDCOM-Datei und Vergleich mit den Orten der näheren Verwandtschaft (nicht wie im Algorithmus mit den Clustern der Quelle generell). Ortsbezeichnungen in den Ländern der inkludierten Mini-GOVs wurden hierbei wie Orte in Deutschland behandelt.
[81] Dieser Schritt wurde im Verlauf der Erarbeitung iterativ durchgeführt, um ungünstige Funktionen des Programmcodes zu entdecken und zu verbessern.

GOV-Daten sind unvollständig (Klassifiziert nach den heutigen Bundesländern: Schleswig-Holstein 12x, Rheinland-Pfalz 7x, Saarland 4x, Sachsen 2x, Baden-Württemberg 2x, Hessen 2x, Bayern 1x, Thüringen 1x, Sachsen-Anhalt 1x)	32
Ausgestaltung des Algorithmus deckt Fall nicht ab	6

Tab. 5: Fehler in der Zuordnung der Provinzen bei der Bezugszeit 1820. [Goldberg 2022]

Es zeigt sich, dass der wesentliche Grund für den hohen Grad der Nicht-Erkennung darin liegt, dass die Objekte außerhalb des relevanten Zielgebiets liegen oder aber die GOV-Daten nicht vollständig sind. Letzteres betrifft vor allem die Orte in den heutigen Bundesländern Schleswig-Holstein, Rheinland-Pfalz und im Saarland. Hochgerechnet etwa 94 Prozent der Fehler liegen als nicht primär im Algorithmus begründet. Daraus ergibt sich eine Trefferquote von 97 Prozent[82] für die gesamte Provinzenzuordnung bezogen auf die GOV-URIs im Betrachtungsgebiet bei denen das GOV vollständig gepflegt ist. Hier zeigt sich, dass das GOV bereits eine breite Datengrundlage enthält und viele Fälle abdecken kann, jedoch in Bezug auf die historischen Verwaltungszugehörigkeiten auch künftig weiter komplettiert werden sollte.

Bei einer Änderung der Bezugszeit auf das Jahr 2020 ergibt sich ein ähnliches Bild (siehe Tabelle 6). Auffällig ist, dass die GOV-Verwaltungszugehörigkeiten hier vollständiger sind. Am deutlichsten ist das bei Schleswig-Holstein. Für die Bezugszeit 2020 ergibt sich eine Trefferquote von 96 Prozent. Die Fehler sind allerdings etwas anders bedingt. Zwar sind sie auch darauf zurückzuführen, dass das GOV unvollständig ist. Das Bundesland ist jedoch (anders als oben) Teil der übergeordneten Objekte; lediglich die Dauer der Zugehörigkeit ist nicht gepflegt. Das tritt bei einzelnen Orten in Bayern und in Schleswig-Holstein um Lübeck herum auf.

Grund der fehlgeschlagenen Klassifizierung	Anzahl
Ort liegt außerhalb der Zielprovinzen (In heutigen Staaten: Niederlande 41x, Polen 22x, Österreich 12x, Tschechien 10x, Frankreich 5x, Schweiz 2x, Dänemark 1x)	93
Ausgestaltung des Algorithmus deckt Fall nicht ab (Klassifiziert nach den heutigen Bundesländern: Schleswig-Holstein 3x, Bayern 3x, Rheinland-Pfalz 1x)	7

Tab. 6: Fehler in der Zuordnung der Provinzen bei der Bezugszeit 2020. [Goldberg 2022]

Insgesamt werden mit dem Algorithmus – unabhängig der beiden getesteten Bezugszeiten – etwa drei von vier Ortsangaben richtig identifiziert und lokalisiert. Werden die Ortsangaben isoliert betrachtet, die einer Provinz zugeordnet werden, so sind hier ebenfalls etwa drei von vier Zuordnungen korrekt. Fast jede identifizierte Ortsangabe kann nachfolgend auch regional geclustert werden.

6. Zusammenfassung

Ob das Neustadt an der Aisch oder in Sachsen, an der Weinstraße oder in Hessen gemeint ist, kann durch den vorgestellten Algorithmus nun ermittelt werden – er trifft eine Entscheidung ohne menschliches Zutun. Grundlage hierfür stellt der Kontext dar, in dem die Bezeichnung des Ortes genannt wird. Die Kontextsensitivität wird dadurch erreicht, dass ein Bezug zu anderen Ortsangaben in der gleichen Quelle hergestellt wird. Die grundlegende Heuristik dahinter besagt: Gemeinsam genannte Orte

[82] [Anzahl gefundener Provinzen] ÷ [Anzahl gefundener Provinzen + (1 – Anteil Fehler) × Anzahl nicht gefundener Provinzen].

liegen beieinander. Durch die Formalisierung und Automatisierung dieser Heuristik ergibt sich eine kontextsensitive Anwendung, die auf verschiedenen wissenschaftlichen Gebieten genutzt werden kann. Der Algorithmus trifft bei Anwendung auf die GEDCOM-Dateien der öffentlich zugänglichen genealogischen Datenbank GEDBAS – nach einer Bereinigung – in drei von vier Fällen eine korrekte Entscheidung. Das ist vor dem Hintergrund ein guter Wert, dass die Daten oftmals eine schlechte Qualität aufweisen, kein Standard in der Benennung praktiziert wird und sie Urbanonyme aus aller Welt enthalten, wenngleich der Schwerpunkt im zentraleuropäischen Raum zu verorten ist.

Die Ortsbezeichnungen werden mithilfe des Kontextes identifiziert, durch die Bestimmung der Koordinaten lokalisiert und anschließend in einer (historischen) Provinz regional geclustert. Für den letzten Aspekt ist eine Bezugszeit zu definieren. Der Algorithmus ist derzeit darauf ausgelegt, Orte innerhalb des deutschsprachigen Raums (ohne Österreich und die Schweiz) in den Grenzen von 1815 bis 1871 sowie ab 1990 zuzuordnen. In diesem Rahmen kann sich auch die Bezugszeit bewegen. Dazu wird das Geschichtliche Orts-Verzeichnis (GOV) genutzt, ein geographisches Lexikon, das stetig erweitert wird. Die Zuordnung zu einer definierten Provinz gelingt in 70 Prozent der Fälle. Werden diejenigen Fehler herausgefiltert, die auf einer unvollständigen Datengrundlage und Orten außerhalb des Betrachtungsgebiets basieren, ergibt sich eine Zuordnungsrate von 96 Prozent.

Zusammengefasst werden drei von vier relevanten Ortsangaben lokalisiert, wovon wiederum drei Viertel korrekt sind. Über 90 Prozent der Orte im Betrachtungsgebiet können nachfolgend regional geclustert werden. Der Algorithmus bietet damit eine geeignete Grundlage, um ihn an verschiedene Anwendungsszenarien anzupassen. Er stellt dadurch ein frei verfügbares Werkzeug insbesondere für wirtschafts- und sozialgeschichtliche Untersuchungen dar. Er kann prädestiniert dafür eingesetzt werden, Fragen der räumlichen Mobilität zu erörtern oder einzelne Datensätze in historischen Provinzen zu klassifizieren. Vor allem der Einsatz in kilometrischen Studien ist vorstellbar. Vorteile ergeben sich insbesondere in der Verarbeitung von Massendaten, wodurch neue Perspektiven für die Wissenschaft eröffnet werden. Er ist besonders dann vorteilhaft, wenn viele Ortsangaben im Kontext erfasst sind. Das ist u. a. der Fall, wenn zahlreiche zusammenhängende Orte ausgewertet und lokalisiert werden müssen. Zur Validierung des Algorithmus wurden zwar genealogische Daten genutzt. Prinzipiell ist der Algorithmus aber bei all solchen Quellen empfehlenswert, bei denen im Kontext weitere Ortsbezeichnungen genannt werden und ein räumlicher Zusammenhang nicht explizit verneint werden kann.[83]

Derzeit deckt der Algorithmus den historischen deutschen Sprachraum ab. Da Informationen des GOVs auch für weitere europäische Länder verfügbar sind, könnten diese samt der dazugehörigen Regionen folgend integriert werden. Der Algorithmus ist zwar auf die deutsche Sprache ausgelegt (z. B. der Einsatz der Kölner Phonetik und die String-Bereinigungsregeln). Dennoch könnte er auf weitere Länder / Sprachen angepasst werden. Auch ist eine Weiterentwicklung des Kontextbezugs denkbar. Ein Beispiel dafür stellen die im Kontext genannten Vor- und Familiennamen dar. Voraussetzung hierfür ist eine vorhergehende Erstellung einer Datenbasis über die zeitliche und räumliche Verbreitung von Namen, auf die sich der Algorithmus beziehen kann.

[83] Der Algorithmus sollte zudem nicht auf antike oder mittelalterliche Quellen angewendet werden, da die hier verwendete Referenzdatenbank, das GOV, keine Daten über solch frühe Strukturen enthält. Darüber hinaus bestehen andere Herausforderungen, die einer gesonderten Betrachtung bedürfen. Die Idee der kontextsensitiven Entscheidungsfindung über geographische Distanzen kann jedoch auch bei solchen Quellen Anwendung finden.

Bibliographische Angaben

Gregory Dominic Abowd / Anind K. Dey / Peter J. Brown / Nigel Davies / Mark Smith / Pete Steggles: Towards a Better Understanding of Context and Context-Awareness. In: Handheld and Ubiquitous Computing. Hg. von Hans-Werner Gellersen. Berlin / Heidelberg 1999. S. 304-307. [Nachweis im GVK]

Jürgen Bähr / Christoph Jentsch / Wolfgang Kuls: Bevölkerungsgeographie. Berlin u. a. 1992. [Nachweis im GVK]

Bertrand Clarke / Ernest Fokoue / Hao Helen Zhang: Principles and Theory for Data Mining and Machine Learning. Dordrecht u. a. 2009. (= Springer Series in Statistics) [Nachweis im GVK]

Martin Ester / Hans-Peter Kriegel / Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings. Second International Conference on Knowledge Discovery & Data Mining. Hg. von Evangelos Simoudis / Jiawei Han / Usama Fayyad (KDD-96: 2,Portland, OR, 02.–04.08.1996) Menlo Park, CA 1996. PDF. [online] [Nachweis im GVK]

Tom Fawcett: An introduction to ROC analysis. In: Pattern Recognition Letters 27 (2006), H. 8, S. 861–874. [Nachweis im GVK]

Computers and thought. Hg. von Edward A. Feigenbaum / Julian Feldman. New York, NY u. a. 1963. [Nachweis im GVK]

Georg Fertig / Christian Schlöder / Rolf Gehrmann / Christina Langfeldt / Ulrich Pfister: Das postmalthusianische Zeitalter: Die Bevölkerungsentwicklung in Deutschland, 1815–1871. In: Vierteljahrschrift für Sozial- und Wirtschaftsgeschichte 105 (2018), H. 1, S. 6–33. [Nachweis im GVK]

Andrew C. Gallagher / Tsuhan Chen: Estimating age, gender, and identity using first name priors. In: IEEE Conference on Computer Vision and Pattern Recognition (2008). DOI: 10.1109/CVPR.2008.4587609 [Nachweis im GVK]

Corry Gellatly: Reconstructing Historical Populations from Genealogical Data Files. In: Population Reconstruction. Hg. von Gerrit Bloothooft / Peter Christen / Kees Mandemakers / Marijn Schraagen. Cham u. a. 2015, S. 111–128. [Nachweis im GVK]

Geolocation Techniques: Principles and Applications. Hg. von Camillo Gentile / Nayef Alsindi / Ronald Raulefs / Carole Teolis. New York, NY u. a. 2013. [Nachweis im GVK]

Ausführliche Auswertung: Vornamen 2018. Hg. von Gesellschaft für deutsche Sprache e. V. In: gfds.de. Wiesbaden u. a. 2019. Pressemitteilung vom 02.05.2019. [online]

Simple Heuristics that make us smart. Hg. von Gerd Gigerenzer / Peter M. Todd. New York, NY u. a. 1999. [Nachweis im GVK]

Jan Michael Goldberg / Katrin Moeller: Automatisierte Extraktion und Lemmatisierung historischer Berufsbezeichnungen in deutschsprachigen Datenbeständen. In: Zeitschrift für digitale Geisteswissenschaften 6 (2021). DOI: 10.17175/2022_002

J. Tuomas Harviainen / Bo-Christer Björk: Genealogy, GEDCOM, and popularity implications. In: Informaatiotutkimus 37 (2018), H. 3, S. 4–14. DOI: 10.23978/inf.76066

Karl Hohensinner: Der Name Mayr / Mair / Mayer / Maier etc. im Oberösterreichischen Familiennamenatlas. In: Familiennamengeographie. Ergebnisse und Perspektiven europäischer Forschung. Hg. von Rita Heuser / Damaris Nübling / Mirjam Schmuck. Berlin u. a. 2011, S. 91–106. [Nachweis im GVK]

Das Historische Ortsverzeichnis von Sachsen. Hg. von Institut für sächsische Geschichte und Volkskunde. In: hov.isgv.de. Dresden 2020. [online]

Jürgen Kocka / Karl Ditt / Josef Mooser / Heinz Reif / Reinhard Schüren: Familie und soziale Plazierung. Studien zum Verhältnis von Familie, sozialer Mobilität und Heiratsverhalten an westfälischen Beispielen im späten 18. und 19. Jahrhundert. Opladen 1980. [Nachweis im GVK]

Vladimir Iosifovič Levenštejn: Binary Codes Capable of Correcting Deletions, Insertations, and Reversals. Soviet Physics /Doklady 10 (1966), H. 8, S. 707–710. [Nachweis im GVK]

Johann-Mattis List: Distanz- und Alignmentanalysen in derhistorischen Linguistik. Düsseldorf 2010. PDF. [online]

Michael Rosemann / Jan Recker: Context-aware Process Design: Exploring the Extrinsic Drivers for Process Flexibility. In: Proceedings of the Workshops and Doctoral Consortium. Hg. von M. Petit / T. Latour. Belgium 2006, S. 149–158. [online]

Wilfried Seibicke: Die Personennamen im Deutschen. Berlin u. a. 1982. [Nachweis im GVK]

Wassiou Olarewaju Sitou: Requirements-Engineering kontextsensitiver Anwendungen. München u. a. 2009. PDF. [online]

Christoph Stöpel (2021a): Geogen v4. In: geogen.stoepel.net. 2005–2021. [online]

Christoph Stöpel (2021b): Geogen Onlinedienst. Allgemeine Informationen / Häufige Fragen (FAQ). In: legacy.stoepel.net/. 2005–2021. [online].

GOV/Webservice. Hg. von Verein für Computergenealogie e. V. In: GenWiki. 2015. Beitrag vom 01.11.2015. [online]

The Historic Gazetteer. Hg. von Verein für Computergenealogie e. V. (2021a) In: Genealogy.net. 2021. [online]

Index of /gov/minigov. Hg. von Verein für Computergenealogie e. V. (2021b) In: Genealogy.net. 2021. [online]

Neustadt in Europe - Overview. Hg. von Working Group Neustadt in Europa. Overview. In: neustadt-in-europa.de. Neustadt an der Weinstraße 2021. [online]

Ivo Zandhuis / Menno den Engelse / Edward Mac Gillavry: Dutch Historical Toponyms in the Semantic Web. In: Population Reconstruction. Hg. von Gerrit Bloothooft / Peter Christen / Kees Mandemakers / Marijn Schraagen.Gerrit Bloothooft et al. Cham u. a. 2015, S. 23–41. [Nachweis im GVK].

Jesper Zedlitz / Norbert Luttenberger: A Survey on Modelling Historical Administrative Information on the Semantic Web. In: International Journal on Advances in Internet Technology 7 (2014), H. 3/4, S. 218–231. [online]

Yong Zheng / Shephalika Shekhar / Alisha Anna Jose / Sunil Kumar Rai: Integrating context-awareness and multi-criteria decision making in educational learning. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. Hg. von Association for Computing Machinery. (SAC '19: 34, Limasol, 08.–12.04.2019) New York, NY 2019, S. 2453–2460. [Nachweis im GVK]

Abbildungs- und Tabellenverzeichnis