Universität Zürich
Geographisches Institut
Abteilung Geographische Informationssysteme
Winterthurerstrasse 190
8057 Zürich

Räumliche Beziehungen in einer Bilddatenbank

Untersuchung der intuitiven Anwendung des räumlichen Relationsbegriffes *near* in Grossbritannien

Diplomarbeit

Fakultätsvertreter: Prof. Dr. Robert Weibel
Betreuung: Dr. Ross Purves, Dr. Alistair Edwardes

David Müller
Friesenbergstrasse 13
8055 Zürich

Zürich, November 2008

Inhaltsverzeichnis

Abbildungsverzeichnis

Zusammenfassung

Die vorliegende Diplomarbeit untersucht die Wahrnehmung des Raumes resp. von Distanzen anhand des ungenauen qualitativen räumlichen Relationsbegriffes *near* [dt.: nahe, in der Nähe von]. Zwei Hauptforschungsfragen lauten wie folgt:

- Wird *near* in verschiedenen Regionen mit unterschiedlichen Distanzen angewendet? Falls ja, was sind die Gründe dafür?
- Ist eine Korrelation ersichtlich zwischen der implizierten Distanz mit *near* und der Grösse einer Stadt, worauf sich *near* bezieht?

In geographischer Hinsicht fokussiert sich diese Arbeit auf Grossbritannien. Es wird untersucht, wie *near* in gewissen Regionen in England, Schottland und Wales eingesetzt wird. Um dies herauszufinden, wird eine Bilddatenbank ausgewertet, die Kommentare über den Bildinhalt von Fotografien, die den geographischen Charakter repräsentieren, enthält. Aus diesen Bildkommentaren werden alle räumlichen Beziehung, welche die Struktur A *near* B aufweisen, extrahiert. A ist die Koordinate des Aufnahmeortes der Fotografie und B ein Toponym, wobei Toponyme sämtliche Einträge auf einer topographischen Landeskarte darstellen. Dies kann daher nicht nur eine Stadt oder ein Dorf, sondern ebenfalls ein Stadtteil, eine Siedlung, ein Gebäude, ein Park, ein Platz, ein Hügel, ein Wald, usw. sein. Das Toponym B aus der Bilddatenbank wird mit einem Gazetteer abgeglichen, um die Koordinate von B zu erhalten. Ein Gazetteer ist ein Register, das alle Toponyme mit den entsprechenden Koordinaten enthält. Nachdem nun auch das Toponym über eine Koordinate verfügt, kann die Distanz zwischen A und B berechnet werden.

Anhand der ersten Fragestellung kann folgendes festgehalten werden: Durchschnittlich wird *near* in Grossbritannien für eine Distanz von 2.28 km verwendet, wobei auf Grund verschiedener Effekte regional unterschiedliche Werte beobachtet werden können. In Schottland zum Beispiel werden im Zusammenhang mit *near* grössere Distanzen verwendet als in Südengland, da im Norden Grossbritanniens eine geringere Toponym-Dichte vorherrscht als in Südengland, wo die Chance daher höher liegt, dass man sich von einem Punkt A auf ein näher liegendes Toponym B

bezieht. Aus diesem Grund wird in Schottland eine grössere Distanz mit *near* im Zusammenhang mit Bildbeschriftungen impliziert als in Südengland - die konzeptionelle Distanz ist auf Grund der Toponym-Dichte regionalbedingt.

Die zweite Fragestellung bringt folgende Erkenntnis hervor: Verwendet man *near* im Zusammenhang mit einem Toponym, das weder eine Stadt noch ein Dorf ist, wird eine mittlere Distanz von 1.94 km assoziiert. Stellt das Toponym ausschliesslich eine Stadt oder ein Dorf dar, ruft dies eine Distanz von 3.51 km hervor. Betrachtet man bloss letzteres Resultat, kann beobachtet werden, dass die Grösse bzw. die Einwohnerzahl einer Stadt wiederum keinen Einfluss auf die Distanz hat. Diese Diplomarbeit hält weitere Effekte fest, weshalb die Perzeption der Distanz in der Verwendung mit *near* in verschiedenen Regionen Grossbritanniens unterschiedlich ist. Die daraus gewonnen Erkenntnisse können weiterentwickelt werden, so dass diese in GIS-Programmen zum Beispiel in Form einer *near*-Funktion implementiert werden können. Die Funktion berechnet, welcher Sektor eines Untersuchungsgebietes als *nahe* zu einem Punkt bezeichnet werden kann.

Abstract

The aim of this diploma thesis is to analyse the perception of space and distance on the basis of the imprecise, qualitative spatial relationship described by the term *near*. There are two main research questions:

- Are there varying perceptions of distances, in different regions, in relation to the use of the term *near*? If so, what are the reasons for this?
- Is there a correlation between the distance implied by *near* and the size of a settlement, which *near* is being used in relation to?

The thesis focuses on the region of Great Britain, investigating under what conditions the term *near* is applied in different regions of England, Scotland and Wales. To consider this, a large image database is used, where each image represents a characteristic view of the position it was taken from. The database contains comments about the picture-contents. It is possible to extract from some of these comments the structure „A *near* B“, where A is the position of the photographer and B a toponym. Such toponyms can be cities, villages, streets, lakes etc.

The answer to the first question shows, that the average for *near* in Great Britain is a distance of 2.28 km (with a standard deviation of 4.89 km), in relation to picture captions. Furthermore, there are several effects which cause different results in different areas of Great Britain. For instance, the results show larger distances for the average of *near* in the north than in the south. The reason for this is the lower density of the toponyms in the north. That means: The chance to relate to a *nearer* toponym in Scotland is smaller than it is in the south of England. This illustrates the influence of the density of toponyms on the conceptualisation of distance (when used in the context of picture captions).

The second question shows that: If *near* is used in combination with a toponym that is not a city or a village, the average distance is 1.94 km. When toponyms are cities or villages, the average of the applied distances is 3.51 km. However, the size of a city or a village does not

influence the distance value significantly. These and other findings described in this work could be implemented in the form of a *near*-tool in GI software. Such an operation could calculate what region can be described as being *near* to a certain point inside an investigation area.

Kapitel 1

Einleitung

1.1 Einführung in die Thematik

Geographische Informationssysteme (GIS) werden von einem immer breiteren Publikum benutzt, sei es in Geowissenschaften, in Raumplanung, in Verkehrsplanung, in Ökologie, in Ökonomie oder Psychologie. Um den Nutzern von GIS einen möglichst intuitiven Zugang zu gewährleisten, muss das kognitive Basiswissen des Menschen über seine Umwelt erfasst und verstanden werden können. Das Verständnis darüber erleichtert die Implementierung dieses Wissens in GIS-Programme (Egenhofer und Mark, 1995). Die vorliegende Arbeit soll einen Beitrag zu räumlichen Datenanalysen leisten, indem die Wahrnehmung des Menschen von Distanzen erforscht und formalisiert wird.

Menschen beschreiben einen Standort im Raum in der Alltagssprache oft mit ungenauen qualitativen räumlichen Relationsbegriffen wie links, rechts, südlich, nördlich oder *nahe* bzw. *in der Nähe* [engl. *near*] (Yao und Thill, 2006). In dieser Diplomarbeit wird der Fokus auf den letzten Begriff „*near*“ gerichtet, da *near* besonders häufig verwendet wird, um einen Standort im Raum zu beschreiben (Yao und Thill, 2007) und somit ein interessanter Forschungsgegenstand ist. Das Beispiel in der Abbildung 1.1 veranschaulicht die Variabilität bzw. die Relativität der ungenauen Bezeichnung *near* - je nach dem, auf welche Stadt referenziert wird. Dabei ist zu beachten, dass *far* nicht zwingend das Gegenteil von *near* darstellen muss. Die Bereiche, die diese beiden Begriffe abdecken, können auch ineinander überfliessen.

Die Datengrundlage dieser Arbeit bildet eine Bilddatenbank, die englischsprachige Kommentare über den Bildinhalt von Fotografien enthält. Sämtliche Resultate dürfen daher nur im Zusammenhang mit Bildkommentaren interpretiert werden. Hinter dieser Datenbank steht die

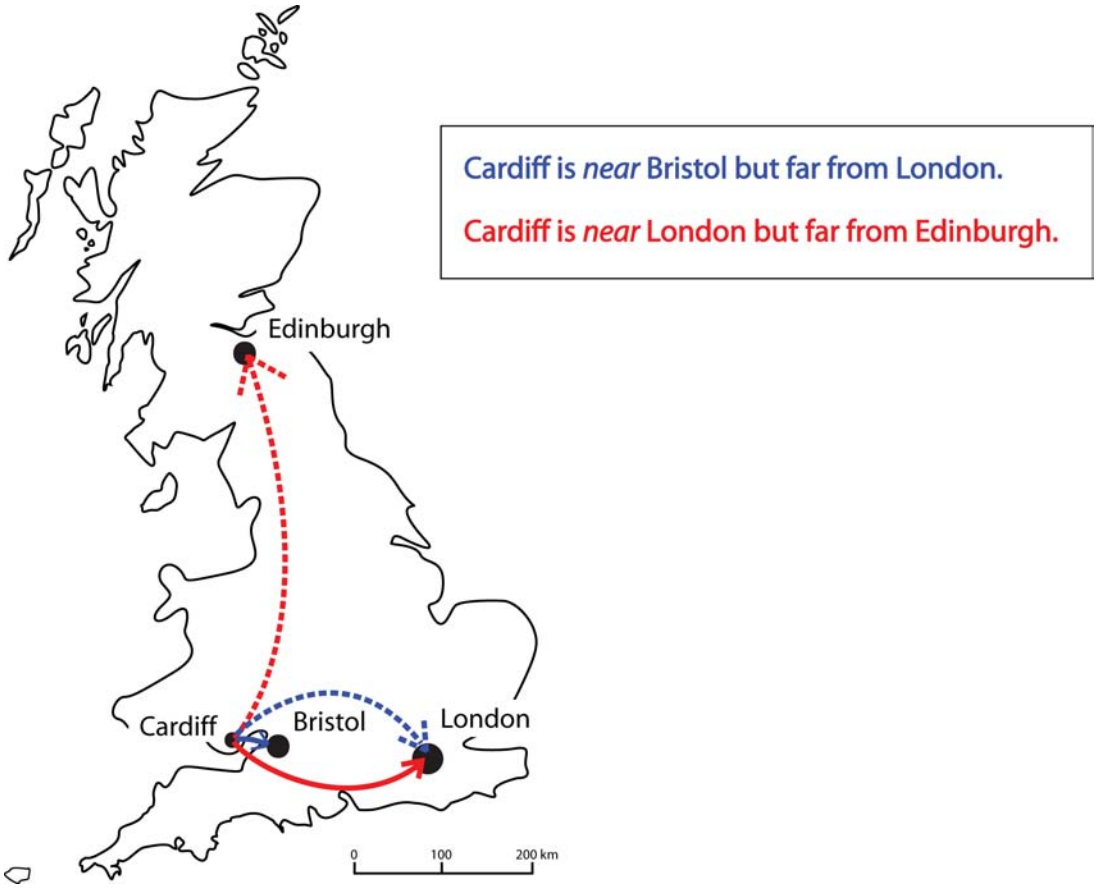


Abbildung 1.1: Interpretation der linguistischen Variable *near*.

Homepage www.GEOGRAPH.org.uk, die zum Ziel hat, jeden Quadratkilometer Grossbritanniens mit Fotografien abzudecken, deren Bildinhalt den geographischen Charakter dieses Gebietes repräsentiert. Die Fotografen (bzw. Teilnehmer) müssen zu jedem Bild ein Kommentar verfassen und das Sujet einer vorgegebenen Bildkategorie (folgend „Imageclass“ genannt) zuordnen. Aus diesen Bildkommentaren werden alle räumlichen Beziehung extrahiert, welche die Struktur A *near* B aufweisen: Etwa jeder 25. Bildkommentar entspricht den Anforderungen, worauf noch 14’861 gültige Beziehungen bzw. Stichproben in der Form A *near* B vorliegen. A ist die Koordinate des Standortes des Fotografen und B stellt ein Toponym dar. Toponyme stellen sämtliche Einträge auf einer Landeskarte im Massstab von 1:50’000 dar. Dies kann daher nicht nur eine Stadt oder ein Dorf, sondern ebenfalls ein Stadtteil, ein Gebäude, ein Berg, ein See usw. sein. Das Toponym B aus der Bilddatenbank wird anschliessend mit einem 50k-Gazetteer abgeglichen, um die Koordinate von B zu erhalten. Ein 50k-Gazetteer ist ein Register, das alle Toponyme (mit den entsprechenden Koordinaten) enthält, die auf einer 1:50’000 topographischen Landeskarte

eingetragen sind.

Nach dieser ersten Extraktion werden die verbleibenden Bildkommentare weiteren Kriterien unterzogen. Viele Toponyme sind identisch geschrieben, haben jedoch eine unterschiedliche geographische Lage. Zum Beispiel ist das Toponym „Abbey Fm“ 50-mal im Gazetteer registriert. Es muss daher überprüft werden, auf welches sich der Teilnehmer tatsächlich bezieht. Würde man diese Mehrdeutigkeit (Ambiguität) nicht berücksichtigen, entstünden riesige Distanzen, die unmöglich im Zusammenhang mit *near* verwendet werden; dies würde eine Verfälschung der Resultate zur Folge haben. Der letzte Schritt besteht darin, die Distanz zwischen den beiden Koordinaten A und B zu berechnen.

Basierend auf bisherigen Forschungsergebnissen werden die folgenden Hauptforschungsfragen vorgestellt, die im Kapitel 2.4 noch eingehender erläutert werden:

- Sind die Distanzen zwischen A und B zufällig entstanden?
- Sind regionale Unterschiede in der Verwendung von *near* zu beobachten? Falls ja, was sind die Gründe dafür?
- Ist eine Korrelation zwischen der Einwohnerzahl und der Distanz zwischen A und B zu beobachten?
- Hat die Verteilung der Toponyme im Gazetteer einen Einfluss auf die Distanzen der Stichproben?

Das Zitat von Robinson (1990, S. 857-872) drückt die Herausforderung aus, unscharfe Begriffe wie *near* zu spezifizieren bzw. zu formalisieren: *„Some spatial relations have an inherent fuzziness about them that is difficult, if not impossible, to specify. One of those spatial relations is near.“*

1.2 Einschränkung der Thematik

Forschungsgegenstand in dieser Diplomarbeit ist die Beziehung A *near* B. Diese Wendung bedeutet eine starke Einschränkung, um einen Standort im Raum zu beschreiben. Theoretisch wäre es möglich, weitere Variablen wie *at, in, on, next to, beside, by, to the left/-right, south, north,* usw. aus der Datenbank von GEOGRAPH (nachfolgend auch GEO-DB genannt) zu extrahieren. Damit würde jedoch die Komplexität der Thematik drastisch ansteigen, was den inhaltlichen und zeitlichen Umfang einer Diplomarbeit sprengen würde.

Es werden fünf Faktoren dahingehend untersucht, ob und wie sich die Wahrnehmung von Distanzen beeinflussen: die (euklidische) Distanz, der Referenzrahmen, die Region, die Toponym-Dichte und die Toponym-Verteilung im Gazetteer. Faktoren wie Reisezeit, Reisekosten, Attraktivität des Zielortes, Aktivität am Zielort usw. (Gahegan, 1995; Hernández et al., 1995; Guesgen, 1999; Yao und Thill, 2005) werden in dieser Arbeit nicht untersucht. In geographischer Hinsicht ist die Arbeit mit England, Schottland und Wales auf Grossbritannien begrenzt.

1.3 Motivation

Forschung über räumliche Relationsbegriffe, die in Form von linguistischen Variablen wie zum Beispiel *near* vorliegen, sind im GIS-Bereich sehr aktuell. Das Ziel besteht darin, qualitative Aussagen so zu erfassen, dass diese in GIS implementiert werden können, um den Benutzern einen möglichst intuitiven Zugang zu gewähren (Egenhofer und Mark, 1995). An der GIS-Abteilung am Geographischen Institut der Universität Zürich läuft momentan ein EU-unterstütztes Projekt namens „TriPod“ (TRI-Partite multimedia Object Description), welches sich unter anderem mit automatischer Bildbeschriftung beschäftigt (Purves et al., 2008; Edwardes et al., 2008). Diese Diplomarbeit erforscht Bildbeschriftungen, die in der Form A *near* B vorliegen, wobei A der Aufnahmeort des Bildes und B ein Toponym ist.

Die Erkenntnisse über die Verwendung von *near* können zu einem Suchalgorithmus weiterentwickelt werden, den man in einer Internetsuchmaschine integrieren könnte; somit weiss die Suchmaschine bei einer Abfrage mit *near*, aus welchem Umkreis die Treffer aufgelistet werden sollen. Erstmals wird untersucht, ob auch das Gazetteer einen Einfluss auf die verwendeten Distanzen im Zusammenhang mit einer qualitativen Variable wie *near* hat. Ist dies der Fall, müssen weitere Untersuchungen, die sich auf das Gazetteer stützen, dieses Verhalten mitberücksichtigen. Der sehr grosse Datensatz von GEOGRAPH, der mit knapp 15'000 gültigen Stichproben ganz Grossbritannien abdeckt, stellt die Gelegenheit dar, Forschungen über die Wahrnehmung von Distanzen anzustellen, die in dieser Dimension noch nie durchgeführt wurden.

1.4 Aufbau der Arbeit

Zu Beginn der Arbeit wird im Kapitel „Hintergrund“ auf bereits existierende Literatur im Bereich „räumliche Beziehungen“ und „Proximity“ eingegangen. Damit soll die Grundlage für das räumliche Denken im Zusammenhang mit *near* erarbeitet werden. Darauf werden die vier Hauptforschungsfragen vorgestellt. Im „Grundlagen“-Teil werden die drei Datenbanken (GEO-DB, Gazetteer, Einwohnertabelle) erläutert, welche die Grundlage der Resultate darstellen. In der „Me-

thodik“ wird auf die Prozessierung der Datensätze eingegangen, zum Beispiel, wie die Datensätze miteinander abgeglichen und somit reicher an Informationen werden. Im selben Kapitel wird die mathematische Grundlage erklärt, worauf sich die Distanzmessung stützt. Dabei werden die verschiedenen Metriken vorgestellt und erklärt, weshalb in dieser Arbeit die euklidische Distanzfunktion gewählt wird.

Das Kapitel „Resultate“ beinhaltet statistische Tests wie der Kolmogorov-Smirnov- oder der Mann-Whitney-U-Test. Letzterer Test wird oft verwendet, um zwei Verteilungen auf ihre Unabhängigkeit zu überprüfen. Es wird ermittelt, wie gross die mittlere Distanz zwischen A und B ist in ganz Grossbritannien und in welchen Regionen welche Distanzen im Zusammenhang mit *near* verwendet werden. Einzelne Gebiete werden miteinander verglichen, wie zum Beispiel Edinburgh mit London. Weiter werden Korrelationen zwischen der Distanz von A und B mit einer allfälligen Einwohnerzahl von B gesucht. Im letzten Teil der Resultate wird die Verteilung der Toponyme im Gazetteer näher betrachtet. Dabei wird die Fragestellung behandelt, ob das Gazetteer die Verteilungen der Stichproben beeinflusst. In der „Diskussion“ wird beispielsweise die Frage behandelt, weshalb sich tatsächlich Parallelen zwischen der Verteilung in der GEO-DB und jener im Gazetteer finden. Es werden diverse Effekte, wie zum Beispiel der „Zentrums-Effekt“ entdeckt, welcher dafür verantwortlich ist, dass keine Grossstadt mit einer Distanz kleiner als 4 km verwendet wird. Den Abschluss bildet das Kapitel „Schlussfolgerungen und Ausblick“. Darin werden die wichtigsten Erkenntnisse zusammengetragen und weiterführende Arbeiten angedacht.

1.5 Definitionen

Einige Fachbegriffe und Fremdwörter werden ebenfalls in den entsprechenden Kapiteln definiert. Die Tabelle 1.1 ist eine Zusammentragung der am häufigsten verwendeten Begriffe in dieser Arbeit:

Tabelle 1.1: Definitionen von Begriffen.

near	dt.: nahe, in der Nähe von
Toponym	alle Einträge auf einer Karte, z.B. Städte, Dörfer, Strassen, Pärke, Seen
Gazetteer	Register, das alle Toponyme enthält
A	Koordinate des Aufnahmeortes
B	Koordinate des Toponyms
Ambiguität	[lat.: ambiguus] Doppel- bzw. Mehrdeutigkeit
Ordnance Survey	Amtsstelle in Grossbritannien, die u.a. Karten herstellt
GIS	Geographische Informationssysteme
Perl	Skriptsprache

Kapitel 2

Hintergrund

2.1 Räumliche Beziehungen

Egenhofer und Mark (1995) beschreiben den Ausdruck „Naive Geography“, wobei es sich um das kognitive Basiswissen handelt, über welches die Menschen von der sie umgebenden Geographie verfügen. Das Ziel von Egenhofer und Mark ist es, dieses Wissen zu verstehen, damit dieses in GIS implementiert werden kann. Daraus erhoffen sie sich, dass ein möglichst intuitiver Zugang zu GIS-Programmen entsteht.

Stevens und Coupe (1978) schreiben, dass sich Menschen räumliche Informationen in einer hierarchischen Anordnung merken. Dabei haben sie miteinbezogen, dass nichträumliche Informationen, wie zum Beispiel die Wichtigkeit eines Objektes, diese hierarchische Abstufung beeinflussen können. Friedmann und Brown (2000) nehmen Stellung zu Stevens und Coupes Idee und denken dabei, dass die Repräsentation von räumlichem Wissen nicht unbedingt hierarchisch sein muss. Sie sagen, dass die räumliche Vorstellung vorwiegend durch nichträumliche Informationen gebildet wird und denken eher an eine kategoriale als an eine hierarchische Einteilung der Informationen. McNamara (1991) spricht von der Möglichkeit einer partiellen hierarchischen Strukturierung des räumlichen Wissens.

Guesgen (1999) schreibt, dass Menschen oft präzise quantitative Information in qualitative Werte wandeln, um den Sinn einer Information besser zu verstehen. Kann man diese Konvertierung nachvollziehen, erhofft er sich, dass man GIS entwickeln kann, die sich besser auch auf qualitative Informationen stützen können. Guesgen stellt daher ein Modell vor, das präzises und unpräzises räumliches Denken in GIS vereinigen kann. Eine weitere Arbeit über diesen Versuch der Integration von qualitativen Informationen stammt von Loerch und Guesgen (1997). Sie

beschäftigen sich u.a. mit der Bildinterpretation, was ihnen geometrische und semantische Erkenntnisse über den Bildinhalt liefert. Worboys (1996) stellt im Paper „Metrics and Topologies for Geographic Space“ einen Vergleich an zwischen einer rechnergestützten räumlichen Repräsentation der Welt und derjenigen, wie sie die Menschen verstehen. Damit bringt er auch die Asymmetrie ins Spiel: Eine Wegstrecke ist bezogen auf jede Metrik symmetrisch, das heisst in beide Richtungen misst eine Strecke gleich viele Kilometer. Misst man jedoch für dieselbe Strecke einen anderen Faktor, zum Beispiel die Reisezeit, dann kann die Richtung eine entscheidende Rolle spielen und eine Asymmetrie darstellen.

Frank (1996) geht davon aus, dass die Menschen vor allem in „large-scale spaces“ denken und sich dabei oft auf Himmelsrichtungen stützen, welches eine qualitative Methode darstellt. Handelt es sich um „small-scale spaces“ werden sinnvollerweise keine Himmelsrichtungen verwendet, wie das folgende Beispiel veranschaulicht: „*Der Kugelschreiber liegt im westlichen Teil des Tisches.*“ Diese Beschreibung würde man nach Pullar und Egenhofer (1988) in die Klasse der „topological relations“ einteilen. Grossräumige Beziehungen, die mit Norden, Osten usw. beschrieben werden, fallen in die Klasse der „direction relations“. Eine mögliche Definition von „large-scale spaces“ wäre:

> *„...a space whose structure is at a significantly larger scale than the observations available at an instant. Thus, to learn the large-scale structure of the space, the traveler must necessarily build a cognitive map of the environment by integrating observations over extended periods of time, inferring spatial structure from perceptions and the effects of actions (Kuipers und Levit, 1990, S. 25-43)“.*

Montello (1993) beschreibt in seinem oft zitierten Paper „Scale and Multiple Psychologies of Space“ vier verschiedene Klassen bzw. Referenzrahmen, in die er den psychologischen Raum einteilt. Diese Raumklassen definiert er aus der Perspektive des Menschen, der sich stets im entsprechenden Raum befindet und nicht aus der Sicht eines externen Beobachters:

- Figural: kleiner als ein Mensch
- Vista: von einem Standort aus sichtbar
- Environmental: Fortbewegung benötigt, um alles zu sehen
- Geographical Space: Karte nötig

Eine ähnliche Einteilung des Raumes unternimmt Fabrikant und Buttenfield (2001), wobei sie sich auf die Semantik in einer räumlichen Datenbank beziehen. Sie definieren die folgenden drei räumlichen Bezugsrahmen (spatial frames of references):

- Geographic Space
- Cognitive Space
- Benediktine Space

Ein Hauptmerkmal des „Geographic Space“ ist das Kontinuum der Skala und die unterschiedliche Granularität, mit der man eine Suchanfrage in einer Datenbank unternehmen kann. Beispiele für verschiedene Granularitäten einer Anfrage wären die Suche nach:

- einem Schlüsselwort innerhalb eines Dokumentes,
- einer Wortkombination innerhalb eines Dokumentes (wie z.B. „A *near* B“),
- einem Dokument in einer Dokumentensammlung.

In weiteren Arbeiten (Worboys, 1996; Hernández, 1994; Hernández et al., 1995; Montello, 1998; Friedmann und Brown, 2000) wurden diverse Definitionen zur Unterteilung des Raumes bzw. des „frames of references“ vorgestellt. Der semantische Vergleich von räumlichen Begriffen muss jedoch in den verschiedenen Kultur- und Sprachregionen differenziert werden. Es kann nicht grundsätzlich davon ausgegangen werden, dass angelsächsische Konzepte mit denjenigen Vorstellung anderer Kulturen übereinstimmen (Frank, 1996; Hernández et al., 1995).

2.2 Nähe und Asymmetrie

Eine der „top five“-Paradoxien der Philosophie beschäftigt sich mit der Frage: „What is a heap of sand?“ [dt.: Sandhaufen] (Sainsbury, 1995). Um nur einige von unendlich vielen Beispielfragen zur Veranschaulichung dieses Paradoxon zu nennen: Ab wann ist ein Hügel ein Berg? Ab wann sind viele Sandkörner ein Sandhaufen? Ab wann ist eine Wasserbewegung eine Welle? Dabei handelt es sich um eine reine Definitionsfrage, wo genau man die Grenze ansetzen soll. Sinngemäss dasselbe Problem stellt sich bei der Fragestellung, ab wann (sprich: ab wievielen Metern) man von einem Punkt nicht mehr *nahe* entfernt ist (Fisher, 2000). In Anspielung an das Beispiel mit dem Haufen hat Fisher (2000, S. 7-18) die Problematik der Asymmetrie treffend formuliert:

> *„Neither the threshold of definite heapness, nor of definite non-heapness, has to be the same when increasing the numbers of grains as when decreasing the number.“*

Erste systematische Erforschungen über die Asymmetrie im räumlichen Denken haben Sadalla und seine Mitarbeiter angestellt (Sadalla et al., 1980). Es wird untersucht, wie sich die Wahrnehmung von Distanzen ändert, wenn man sich von einem gut bekannten Objekt (reference

point) auf ein weniger bekanntes Objekt (nonreference point) bezieht und umgekehrt. Sadalla et al. (1980, S. 516-528) hat das Ergebnis seiner Studie wie folgt zusammengefasst:

> *„...the cognitive distance between reference points and nonreference points is asymmetrical; nonreference points were judged nearer to reference points than were reference points to nonreference points...“*

Dieses Ergebnis scheint nachvollziehbar, wenn man bedenkt, dass einem die Distanz „Stadtzentrum-Vorort“ aus der Perspektive des Stadtzentrums weiter scheint als die Distanz „Vorort-Stadtzentrum“. Weitere Arbeiten über die Asymmetrie im geographischen Raum stammen von Egenhofer und Mark (1995); Worboys (1996). Die Wahrnehmung der Distanz bzw. der *Nähe* kann auf die folgenden Faktoren zurückgeführt werden (Gahegan, 1995; Hernández et al., 1995; Guesgen, 1999; Yao und Thill, 2005):

- Kontext/Bezugsrahmen/Skala
- Distanz
- Aktivität am Zielort
- Zeitbudget
- Erreichbarkeit
- Transportmittel
- Reisekosten
- Attraktivität und Bekanntheit des Zielortes

Zur Veranschaulichung des ersten Punktes sei auf das Beispiel in der Abbildung 1.1 im Kapitel 1.1 hingewiesen, wobei Cardiff einerseits nahe bei London liegt und andererseits weit entfernt von London ist, je nach Kontext: *Cardiff is near Bristol but far from London. Cardiff is near London but far from Edinburgh.* Während beim Lesen dieser beiden Sätze automatisch „mental maps“ gebildet werden, wird dem ersten Satz ein kleinerer Referenzrahmen zugeordnet als dem zweiten Satz, dessen Ausdehnung praktisch ganz Grossbritannien umfasst. Es kann beobachtet werden, dass die Wahrnehmung von *near* als eine Funktion der Skala interpretiert werden kann. Kennt man diesen Bezugsrahmen ist es möglich, die linguistische Variable *near* bzw. *far* mit einer „fuzzy membership function“ zu quantifizieren, wobei zum Beispiel 0 die grösstmöglichste Separation und 1 die maximale *Nähe* bedeutet (Gahegan, 1995). Guesgen (1999) dagegen erwähnt, dass der Hauptfaktor der Wahrnehmung der Distanz die absolute Distanz darstellt.

Guesgen (2002) hat in Anlehnung an den von Samet (1990) entwickelten „nearest neighbor searching algorithm“ einen Suchalgorithmus entwickelt, um das nächstliegendste Objekt zu finden. Dabei setzt er die Regel der Transitivität zwischen verschiedenen Objekten voraus, um einen „fuzzy membership grade“ von verschiedenen Objekten zu berechnen. Um die Regel der Transitivität zu erläutern, soll das folgende Beispiel genannt werden (Guesgen, 2002, S. 265-270):

If B is closer to A than C is to A and
if C is closer to A than D is to A,
then B is closer to A than D is to A.

Setzt man zusätzlich eine „symmetry of proximity“ voraus, was bedeutet, dass die euklidische Distanz A zu B gleich gross sein muss wie die Distanz B zu A, können die Beziehungen noch schärfer erfasst werden. Die präziseste Stufe stellt die quantitative Distanzmessung dar. Die hierarchisch tiefst gelegene und breiteste Stufe wird durch den allgemeinen Fall repräsentiert (siehe Abb. 2.1) („General“). In dieser Klasse können keine Annahmen über Beziehungen mit *Proximity* getroffen werden. Die zweite Stufe verlangt Transitivität, wie dies im obigen Beispiel erläutert wurde („Transitive“) - werden dabei symmetrische Eigenschaften miteinbezogen, gelangt man die nächste Klasse („Symmetrie“). Die euklidische Distanzmessung stellt als die letzte Stufe die präziseste Messart dar („Distance“).

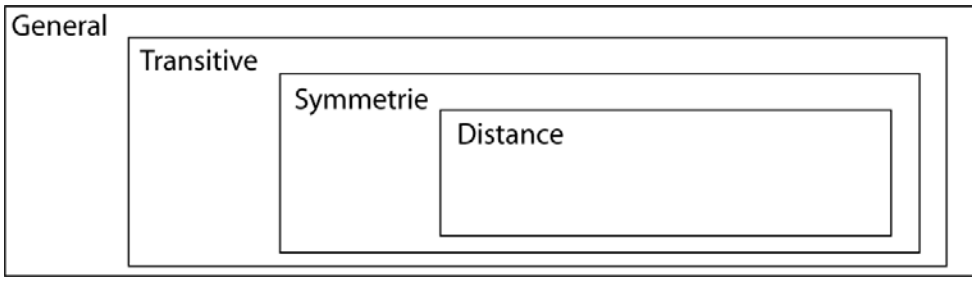


Abbildung 2.1: „Classes of Proximities“ (Guesgen, 2002).

Im Paper „Membership Functions for Spatial Proximity“ von Brennan und Martin (2006) wird kritisiert, dass die bisher erarbeiteten „membership functions“ nie mit einem Experiment auf deren Vorzüge abgeschätzt wurden. Sie denken, dass man mit Experimenten nicht nur objektive Evaluationen durchführen könne, sondern dass auch eine Weiterentwicklung der „fuzzy logic“ resp. der „membership functions“ entstünde. Für eine „membership function“ schlagen Sie eine „fuzzy intersection“ zwischen „relative distance metrics“ und „absolute distance metrics“ vor im Gegensatz zu Gahegan (1995), der eine „fuzzy union“ der beiden Metriken empfiehlt. Robinson (1990) stellt in seinem Paper „Interactive Machine Acquisition of a Fuzzy Spatial Relation“ ein

Programm vor, das versucht *near* zu formalisieren, in dem der Benutzer systematisch abgefragt wird.

Worboys (1996) beschäftigt sich mit dem Faktor „Kontext“ mit dem Ziel, aus einer rein metrischen Skala eine geographische Skala zu generieren, welche asymmetrische Eigenschaften berücksichtigt. Es gibt verschiedenste Ansätze, mit welchen Parametern man den Kontext erfassen könnte. Worboys konzentriert sich bei seinen Untersuchungen auf die 48 populationsreichsten Städte Grossbritanniens und versucht anhand der gegenseitigen Entfernungen der Zentren eine geographische Skala zu erstellen. Er misst von jeder einzelnen Stadt die Distanz zu den umliegenden Städten und berechnet das arithmetische Mittel. Damit kann er die sogenannte „relative Distanz“ (*reldis*) berechnen. Es werden nun die Distanzen zwischen den Städten A und B durch den erwähnten Mittelwert von A dividiert. Die relativen Distanzen sind umso kleiner, je zentraler A liegt, da zentral gelegene Städte im Mittel geringe Distanzen zu den umliegenden Städten aufweisen. Die relative Distanz von A zu B ist demnach ungleich der relativen Distanz von B zu A. Im folgenden Beispiel stellt A die im Nordosten Schottlands gelegene Stadt Aberdeen und B die in der Mitte Englands gelegene Stadt Birmingham dar: Gemäss der oben beschriebenen Berechnung ist $reldis(Aberdeen, Birmingham) = 1.1$ und $reldis(Birmingham, Aberdeen) = 2.6$. Dies erscheint plausibel, da aus der Perspektive von Birmingham, welches eng von vielen Städten umgeben ist, das eher isolierte Aberdeen als relativ weit weg erscheint. Dies unterstreicht die asymmetrische Eigenschaft des geographischen Raumes.

2.2.1 Zwei Experimente von Worboys

Worboys (2001) beschreibt in seinem Paper „Nearness relations in environmental space“ ein Experiment, das er an der Keele University in Stanfordshire durchführte. Die Idee dieses Versuches, die Vagheit der Wahrnehmung von *Proximity* empirisch zu erfassen, hat Worboys aus der Arbeit von Bonini et al. (1999) abgeleitet. Das 2.43 km^2 grosse Universitätsgelände besteht aus 23 markanten Gebäuden, die relativ gleichmässig über den Campus verteilt sind, wobei der Standort der Bibliothek (Library) ungefähr das Zentrum bildet. Am Experiment nehmen 22 Studierende teil, welche in eine Truth- und eine Falsitygroup eingeteilt werden. Alle Teilnehmer der beiden Gruppen erhalten einen Fragebogen, auf dem alle Gebäude aufgelistet sind ausser der „Library“. Die Truthgroup muss neben jedem Gebäude ein Häckchen platzieren, sofern sie dieses Gebäude als *nahe* zur „Library“ empfinden. Die Falsitygroup wird gebeten, dasselbe zu unternehmen, jedoch für jene Gebäude, die sie als *nicht nahe* von der „Library“ empfinden. Die Abbildung 2.2 zeigt das Resultat.

Place	Library: Truth group	Library: Falsity group	Place	Truth group	Falsity group
24 hour Reception	4	4	Holly Cross	1	11
Academic Affairs	5	2	Horwood Hall	4	10
Barnes Hall	0	11	Keele Hall	8	2
Biological Sciences	5	4	Lakes	1	11
Chancellors Building	4	6	Leisure Centre	0	11
Chapel	10	0	Library	11	0
Chemistry	4	6	Lindsay Hall	2	8
Clock House	4	6	Observatory	0	11
Computer Science	1	10	Physics	5	5
Earth Sciences	7	0	Students Union	10	0
Health Centre	1	11	Visual Arts	1	10

Abbildung 2.2: Resultat aus dem „Library-Experiment“ (Worboys, 2001).

Die Stichproben werden mit einem χ^2-Test mit einem Signifikanzniveau von 0.001 auf ihre Unterschiedlichkeit überprüft, was dazu führt, dass 11 Gebäude weder mit *near* noch mit *not near* eindeutig versehen werden können. Dies kann aus zwei Gründen sein: Entweder liegt ein „Truth Gap“ oder ein „Truth Glut“ vor. Ersteres bedeutet, dass weder viele *nears* noch viele *not nears* vorliegen. Zweiteres trifft zu, wenn von beiden Ausprägungen viele vorliegen. Um den Grad des Konfliktes zu beurteilen, verwendet Worboys die „four-valued logic“ nach Belnap (1977). Mit dieser Gewichtungsmethode kann er für jedes Gebäude eine konzeptionelle Distanz zur „Library“ erzeugen.

Das analoge Experiment hat Worboys ein zweites Mal durchgeführt; diesmal wird jedoch untersucht, ob die Probanden die „Lindsay Hall“ als *nahe* bzw. *nicht nahe* einstufen (Worboys et al., 2004). Die „Lindsay Hall“ ist ein weiteres Gebäude auf dem Universitätscampus. Trägt man bei beiden Experimenten die konzeptionelle gegenüber der euklidischen Distanz ab, erhält man das Diagramm in Abbildung 2.3. Interessant ist, dass die konzeptionelle Distanz von der „Library“ zur „Lindsay Hall“ grösser ist als umgekehrt, obwohl die euklidische Distanz gleich ist. Die „Lindsay Hall“ befindet sich in der Peripherie und die „Library“ im Zentrum des Untersuchungsgebietes. Dies bekräftigt die Vermutung, dass man im Zusammenhang mit zentraler gelegenen Objekten kürzere Distanzen assoziiert. Es kann dasselbe Verhalten festgestellt werden, wie im oben erwähnten Beispiel mit den relativen Distanzen zwischen Birmingham und Aberdeen.

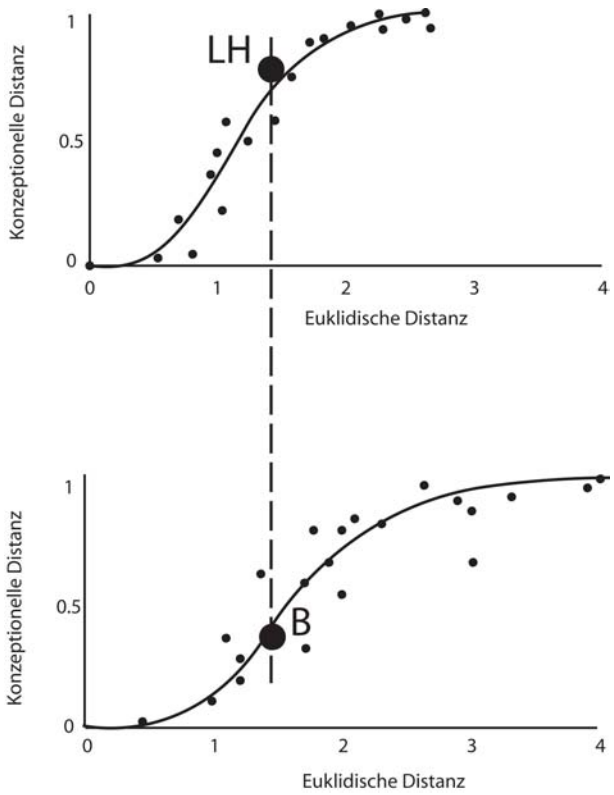


Abbildung 2.3: Oben: Konzeptionelle Distanz zu Gebäuden von der „Library“ aus betrachtet, abgetragen gegenüber der euklidischen Distanz (LH: „Lindsay Hall“).
Unten: Konzeptionelle Distanz zu Gebäuden von der „Lindsay Hall“ aus betrachtet, abgetragen gegenüber der euklidischen Distanz (B: „Library“) (Worboys, 2001).

2.3 Forschungslücke

Es wurde bisher viel über die Wahrnehmung von Distanz bzw. von *Nähe* geforscht. Bei beiden Experimenten nahmen jedoch lediglich eine geringe Anzahl an Probanden teil, bei den Experimenten von Worboys (2001); Worboys et al. (2004) waren es zum Beispiel 22. Die vorliegende Arbeit eröffnet in dieser Hinsicht eine neue Dimension – die zugrunde liegende Datenbank wurde durch 5’345 Teilnehmer erzeugt. Insgesamt enthält die Datenbank 362’308 Bilder bzw. Bildkommentare. Die Resultate in dieser Arbeit stützen sich dadurch mehrheitlich auf relativ grosse Stichproben, was die Grundlage für statistisch gültige Aussagen ist. Die Teilnehmer erstellten Datensätze aus ganz Grossbritannien, womit der Fokus auf einem Untersuchungsgebiet liegt, das noch nie auf die verwendeten Distanzen im Zusammenhang mit *near* untersucht wurde. Die Resultate in dieser Arbeit soll auch einen Beitrag zu Forschungen über Bildabfragen sein. Im Paper von Purves et al. (2008) wird beschrieben, wie Bildabfragen, die den Textabfragen bezüglich

Qualität zur Zeit noch nachstehen, verbessert werden können. Je klarer bzw. eingeschränkter die Struktur eines Bildes ist, desto besser funktioniert die bildinhaltsbasierte Abfrage. Eine gute Anwendung in einer unkontrollierten Bildersammlung, wie dies zum Beispiel auf www.Flickr.com der Fall ist, wäre daher momentan nicht realisierbar.

2.4 Forschungsfragen

Alle Forschungsfragen drehen sich um die Beziehung A *near* B, die aus der Bilddatenbank von GEOGRAPH extrahiert wurde. A stellt dabei die Koordinate des Aufnahmeortes dar und B (nach Abgleichung mit dem Gazetteer) jene des Zielobjektes, das ein Toponym ist. Daher kann man die Beziehung A *near* B als folgende Kombination sehen: *Koordinate des Aufnahmeortes + near + Koordinate des Toponyms.*

2.4.1 Sind die Distanzen zwischen A und B zufällig entstanden?

Im Methodikteil wird erläutert, wie die Distanzen im Datensatz von GEOGRAPH ausgewertet werden. Vorweg soll hier erwähnt werden, dass die Koordinaten der Stichproben 1 km^2 grosse quadratische Kacheln repräsentieren, weshalb die Auswertungen in einem Raster vorgenommen werden. Es wird untersucht, ob die Verteilung der Stichproben im Zusammenhang mit der Verwendung von *near* einerseits und eine zufällige Verteilung in einem Raster andererseits unabhängig voneinander sind. Falls die Verteilung der Stichproben signifikant ähnlich ist mit einer zufälligen Verteilung, würde dies bedeuten, dass die Datenbank von GEOGRAPH willkürlich zu Stande gekommen ist und somit uninteressant wäre.

2.4.2 Sind regionale Unterschiede in der Anwendung von *near* erkennbar?

Es wird untersucht, ob bei der Anwendung von *near* regionale Unterschiede erkennbar sind. Wie stark unterscheiden sich die Beziehungen zwischen dem Schottischen Hochland und dem Englischen Flachland? Wird *near* im Norden wie auch im Süden Grossbritanniens gleich verwendet? Entstehen unterschiedliche Distanzen bei der Auswertung der beiden Hauptstädte Edinburgh und London? Auf Grund dieser Fragestellungen soll ein Bild entstehen, in welchen Regionen grössere bzw. kleinere Distanzen als in anderen Gebieten verwendet werden.

2.4.3 Entsteht eine Korrelation mit der Einwohnerzahl von B und der Distanz zwischen A und B?

All jene Toponyme, welche in der Einwohnertabelle Grossbritanniens auftauchen und somit ein Dorf oder eine Stadt sind, werden dahingehend untersucht, ob zwischen der verwendeten Distanz

im Zusammenhang mit *near* und der Einwohnerzahl des Toponyms eine Korrelation zu erkennen ist. Könnte es sein, dass im Zusammenhang mit grossen respektive mit einwohnerreichen Städten auch grössere Distanzen verwendet werden? Weiter interessiert die Frage, ob diese Beziehungen gegenüber den restlichen Beziehungen (bei denen die Toponyme keine Dörfer oder Städte sind) kleinere Distanzen aufweisen.

2.4.4 Wie sind die Toponyme im Gazetteer verteilt?

Sind im Gazetteer regionale Differenzen bezüglich der Verteilung der Toponyme zu erkennen? Könnte es sein, dass sich allfällige regionale Unterschiede in GEOGRAPH und im Gazetteer ähnlich verhalten? Dabei stellt sich die Frage, ob regionale Unterschiede in GEOGRAPH durch eine Beeinflussung der Verteilung im Gazetteer und nicht etwa durch regional unterschiedliche Wahrnehmungen von der Distanz der Fotografen entstehen.

Kapitel 3

Grundlagen

3.1 Datenbank von GEOGRAPH

Hinter der Bilddatenbank steht die Homepage www.GEOGRAPH.org.uk, welche vor rund dreieinhalb Jahren in Grossbritannien vom englischen Geographen Gary Rogers gegründet wurde. Der Slogan von GEOGRAPH steht für die Idee, welche dahinter steckt: „Photograph every grid square!“ Das Ziel ist es, die charakteristische Geographie oder ein charakteristisches Merkmal jedes Quadratkilometers von Grossbritannien mit mindestens einer Fotografie zu repräsentieren. Teilnehmen darf, wer sich auf der Homepage registrieren lässt und die gestellten Anforderungen erfüllt. Auf den Fotografien dürfen zum Beispiel keine Portraits von Personen abgebildet sein, da dies nicht die lokale Geographie repräsentieren würden. Die User müssen jedes Bild mit einem kurzen Titel sowie mit einem Kommentar versehen. Die Tatsache, dass der Inhalt von GEOGRAPH von über 30 Personen moderiert wird, ist eine enorme Stärke der dabei entstandenen Datenbank, da diese somit klare Richtlinien einhält. Als Gegenbeispiel würde an dieser Stelle das Online Photo Portal www.flickr.com dienen, da dort die Benutzer Bilder aller Art mit irgendwelchen Kommentaren platzieren dürfen. Wie nun die Umgebung einer hochgeladenen Fotografie auf der Homepage von GEOGRAPH schlussendlich aussieht, ist in der Abbildung 3.1 einzusehen.

In der Abbildung 3.1 erkennt man links oben die Koordinate (NJ8437) vom Aufnahmeort des Fotografen gefolgt vom Titel. Diese Koordinate entspricht einem Quadrat mit einer Fläche von 1 km^2. Das ist das vorgegebene Format der Koordinaten A und B in der Datenbank. Unterhalb dieses Titels wird eine anhand der Koordinaten automatisch generierte Beschreibung des Standortes angebracht. Dies interessiert jedoch nicht, da in in dieser Arbeit die intuitive Anwendung des Begriffes *near* den Forschungsgegenstand darstellt und nicht eine automatisch generierte Zuordnung einer Koordinate. Unterhalb des Bildes ist der Kommentar, der zusammen mit der

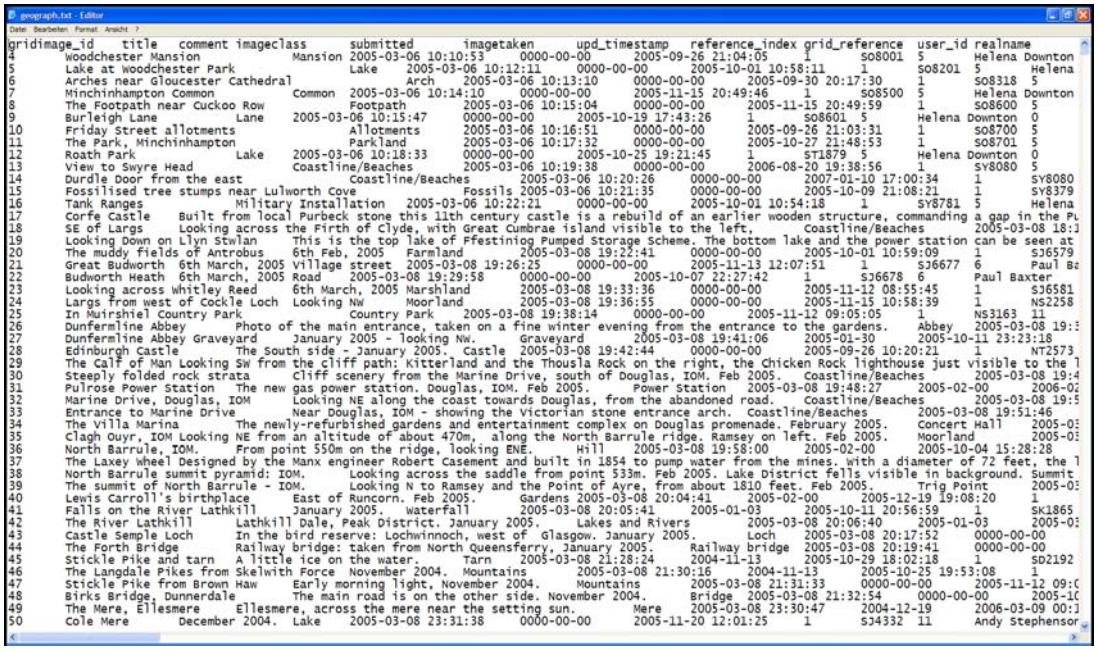
Abbildung 3.1: Beispiel einer Fotografie auf der Homepage von www.GEOGRAPH.uk.org. ©GEOGRAPH

Koordinate die Beziehung A *near* B liefert (NJ8437 near Blackbriggs). Scrollt man die Seite nach unten, dann erscheinen weitere Informationen, wie zum Beispiel der Name des Fotografen, das Aufnahmedatum, die Bildkategorie, das Übermittlungsdatum und die Himmelsrichtung, in welche der Fotoapparat bei der Aufnahme gerichtet war. Die Himmelsrichtung wurde jedoch leider nur bei einer Minderheit aller Aufnahmen vermerkt. Daher stellt dieses Element keine gute Grundlage für allfällige Untersuchungen dar. Nun richtet sich der Fokus dieser Arbeit auf die Titel und Kommentare der Fotografien und auf die Koordinate des Aufnahmeortes. Mit diesen Angaben werden dann weitere Informationen abgeleitet, dazu mehr im Methodikteil.

Bis zum Zeitpunkt dieser Diplomarbeit haben 5'345 Personen mit den oben erwähnten Informationen eine Datenbank generiert. Die meisten Fotografen haben gleich mehrere Bilder übermittelt, so dass die gewaltige Menge von 362'308 Fotografien zusammen kam. Bei der Fläche Grossbritanniens von 330'174 km^2 ergibt dies im Durchschnitt rund 1.1 Bilder pro Quadratkilometer. Da jedoch keine Gleichverteilung über die Fläche herrscht, sind dennoch 46% aller 1 km^2-Gridkacheln noch ohne Bild.

Im Bildkommentar wird oft die geographische Lage des Sujets beschrieben. Es gibt unzählige Arten, wie man über einen räumlichen Relationsbegriff die Lage eines Objektes beschreiben will. Dabei wurde der Relationsbegriff *near* 33'363 mal verwendet, sprich: In etwa jedes zehnte Bild wurde mit *near* beschrieben. Dabei ist ausdrücklich zu bemerken, dass diese Beziehungen

intuitiv zu Stande gekommen sind. Die Fotografen bzw. die Teilnehmer wurden nicht gezwungen, das Wort *near* zu verwenden, sondern haben dies freiwillig bzw., wie gesagt, intuitiv verwendet. Die Fotografen wussten indes auch nicht, dass ihre Kommentare später einer statistischen Analyse unterzogen werden. Informationen, die auf diese im Prinzip inoffizielle Art und Weise zu Stande kommen, werden im Paper von Purves und Edwardes (2008) als „Volunteered Geographic Information“ bezeichnet. Die Kriterien, um aus diesen 33'363 Datensätze die für diese Auswertung relevanten Kommentare auszuwählen, werden im Kapitel 4.1 behandelt. Am schluss dieses Selektionsverfahren bleiben noch 14'861 gültige Bildkommentare übrig. Die Datenbank, die alle Informationen sämtlicher Fotografien enthält, besteht aus einer Textdatei. Eine Zeile entspricht einem Datensatz, der tabulatoren-getrennt Informationen zu einer Fotografie enthält. Die Abbildung 3.2 liefert einen Einblick in die Datenbank, wie sie unverarbeitet von GEOGRAPH geliefert wird.



```
gridimage_id   title   comment imageclass      submitted       imagetaken      upd_timestamp   reference_index grid_reference  user_id realname
4       Woodchester Mansion             Mansion 2005-03-06 10:10:53     0000-00-00      2005-09-26 21:04:05     1       SO8001  5       Helena Downton
5       Lake at Woodchester Park                Lake    2005-03-06 10:12:11     0000-00-00      2005-10-01 10:58:11     1       SO8201  5       Helena
6       Arches near Gloucester Cathedral                Arch    2005-03-06 10:13:10     0000-00-00      2005-09-30 20:17:30     1       SO8318  5
7       Minchinhampton Common           Common  2005-03-06 10:14:10     0000-00-00      2005-11-15 20:49:46     1       SO8500  5       Helena Downton
8       The Footpath near Cuckoo Row            Footpath        2005-03-06 10:15:04     0000-00-00      2005-11-15 20:49:59     1       SO8600  5
9       Burleigh Lane           Lane    2005-03-06 10:15:47     0000-00-00      2005-10-19 17:43:26     1       SO8601  5       Helena Downton  0
10      Friday Street allotments                Allotments      2005-03-06 10:16:51     0000-00-00      2005-09-26 21:03:31     1       SO8700  5
11      The Park, Minchinhampton                Parkland        2005-03-06 10:17:32     0000-00-00      2005-10-27 21:48:53     1       SO8701  5
12      Roath Park              Lake    2005-03-06 10:18:33     0000-00-00      2005-10-25 19:21:45     1       ST1879  5       Helena Downton  0
13      View to Swyre Head              Coastline/Beaches       2005-03-06 10:19:38     0000-00-00      2006-08-20 19:38:56     1       SY8080  5
14      Durdle Door from the east               Coastline/Beaches       2005-03-06 10:20:26     0000-00-00      2007-01-10 17:00:34     1       SY8080
15      Fossilised tree stumps near Lulworth Cove               Fossils 2005-03-06 10:21:35     0000-00-00      2005-10-09 21:08:21     1       SY8379
16      Tank Ranges             Military Installation   2005-03-06 10:22:21     0000-00-00      2005-10-01 10:54:18     1       SY8781  5       Helena
17      Corfe Castle    Built from local Purbeck stone this 11th century castle is a rebuild of an earlier wooden structure, commanding a gap in the Pu
18      SE of Largs     Looking across the Firth of Clyde, with Great Cumbrae island visible to the left,       Coastline/Beaches       2005-03-08 18:1
19      Looking Down on Llyn Stwlan     This is the top lake of Ffestiniog Pumped Storage Scheme. The bottom lake and the power station can be seen at
20      The muddy fields of Antrobus    6th Feb, 2005   Farmland        2005-03-08 19:22:41     0000-00-00      2005-10-01 10:59:09     1       SJ6579
21      Great Budworth  6th March, 2005 Village Street  2005-03-08 19:26:25     0000-00-00      2005-11-13 12:07:51     1       SJ6677  6       Paul Ba
22      Budworth Heath  6th March, 2005 Road    2005-03-08 19:29:58     0000-00-00      2005-10-07 22:27:42     1       SJ6678  6       Paul Baxter
23      Looking across Whitley Reed     6th March, 2005 Marshland       2005-03-08 19:33:36     0000-00-00      2005-11-12 08:55:45     1       SJ6581
24      Largs from west of Cockle Loch  Looking NW      Moorland        2005-03-08 19:36:55     0000-00-00      2005-11-15 10:58:39     1       NS2258
25      In Muirshiel Country Park               Country Park    2005-03-08 19:38:14     0000-00-00      2005-11-12 09:05:05     1       NS3163  11
26      Dunfermline Abbey       Photo of the main entrance, taken on a fine winter evening from the entrance to the gardens.    Abbey   2005-03-08 19:1
27      Dunfermline Abbey Graveyard     January 2005 - looking NW.      Graveyard       2005-03-08 19:41:06     2005-01-30      2005-10-11 23:23:18
28      Edinburgh Castle        The South side - January 2005.  Castle  2005-03-08 19:42:44     0000-00-00      2005-09-26 10:20:21     1       NT2573
29      The Calf of Man Looking SW from the cliff path: Kitterland and the Thousla Rock on the right, the Chicken Rock lighthouse just visible to the l
30      Steeply folded rock strata      Cliff scenery from the Marine Drive, south of Douglas, IOM. Feb 2005.   Coastline/Beaches       2005-03-08 19:4
31      Pulrose Power Station   The new gas power station. Douglas, IOM. Feb 2005.      Power Station   2005-03-08 19:48:27     2005-02-00      2006-02
32      Marine Drive, Douglas, IOM      Looking NE along the coast towards Douglas, from the abandoned road.    Coastline/Beaches       2005-03-08 19:5
33      Entrance to Marine Drive        Near Douglas, IOM - showing the Victorian stone entrance arch.  Coastline/Beaches       2005-03-08 19:51:46
34      The Villa Marina        The newly-refurbished gardens and entertainment complex on Douglas promenade. February 2005.    Concert Hall    2005-03
35      Clagh Ouyr, IOM Looking NE from an altitude of about 470m,  along the North Barrule ridge. Ramsey on left. Feb 2005.    Moorland        2005-03
36      North Barrule, IOM.     From point 550m on the ridge, looking ENE.      Hill    2005-03-08 19:58:00     2005-02-00      2005-10-04 15:28:28
37      The Laxey Wheel Designed by the Manx engineer Robert Casement and built in 1854 to pump water from the mines, with a diameter of 72 feet, the l
38      North Barrule summit pyramid: IOM.      Looking across the saddle from point 533m. Feb 2005. Lake District fells visible in background. Summit
39      The summit of North Barrule - IOM.      Looking N to Ramsey and the Point of Ayre, from about 1810 feet. Feb 2005.      Trig Point     2005-03
40      Lewis Carroll's birthplace      East of Runcorn. Feb 2005.      Gardens 2005-03-08 20:04:41     2005-02-00      2005-12-19 19:08:20     1
41      Falls on the River Lathkill     January 2005.   Waterfall       2005-03-08 20:05:41     2005-01-03      2005-10-11 20:56:59     1       SK1865
42      The River Lathkill      Lathkill Dale, Peak District. January 2005.     Lakes and Rivers        2005-03-08 20:06:40     2005-01-03      2005-03
43      Castle Semple Loch      In the bird reserve: Lochwinnoch, west of  Glasgow. January 2005.      Loch    2005-03-08 20:17:52     0000-00-00
44      The Forth Bridge        Railway bridge: taken from North Queensferry. January 2005.     Railway bridge  2005-03-08 20:19:41     0000-00-00
45      Stickle Pike and tarn   A little ice on the water.      Tarn    2005-03-08 21:28:24     2004-11-13      2005-10-29 18:02:18     1       SD2192
46      The Langdale Pikes from Skelwith Force  November 2004.  Mountains       2005-03-08 21:30:16     2004-11-13      2005-10-25 19:53:08     1
47      Stickle Pike from Brown Haw     Early morning light, November 2004.     Mountains       2005-03-08 21:31:33     0000-00-00      2005-11-12 09:0
48      Birks Bridge, Dunnerdale        The main road is on the other side. November 2004.      Bridge  2005-03-08 21:32:54     0000-00-00      2005-10
49      The Mere, Ellesmere     Ellesmere, across the mere near the setting sun.        Mere    2005-03-08 23:30:47     2004-12-19      2006-03-09 00:1
50      Cole Mere       December 2004.  Lake    2005-03-08 23:31:38     0000-00-00      2005-11-20 12:01:25     1       SJ4332  11      Andy Stephenson
```

Abbildung 3.2: Ausschnitt aus der Datenbank von GEOGRAPH. ©GEOGRAPH

3.2 Ordnance Survey - Great Britain's national mapping agency

Das Koordinatensystem, auf welches sich der zu untersuchende Datensatz stützt, wurde von der Ordnance Survey Great Britain (OSGB), einer Amtsstelle der Regierung des Vereinigten Königreichs, definiert. Wie der Name bereits vermuten lässt (Ordnance heisst auf deutsch Artillerie,

Survey heisst Erkundung), wurde diese Abteilung mit der Absicht gegründet, strategische Karten für das Militär zu generieren. Die ersten Karten wurden von der OSGB bereits im 18. Jahrhundert angefertigt.

Die Abbildung 3.3 veranschaulicht die Systematik des Koordinatensystems. Jeder 100x100 Kilometer Kachel wurde eine Kombination von zwei Buchstaben aus dem römischen Alphabet zugeordnet. Um eine feinere Auflösung des Rasters zu erhalten, werden nach den beiden Buchstaben weitere vier Ziffern angefügt, wobei die ersten beiden Ziffern die x- und die zweiten die y-Achse in Kilometern beschreiben. Ein Beispiel einer Koordinate wäre TQ3181, welche mit 1 km^2 Grösse das Zentrum von London definiert.

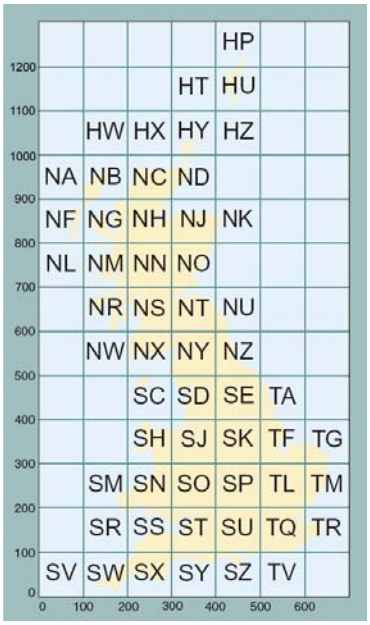


Abbildung 3.3: 100x100 km OSGB Kacheln in Grossbritannien (Ordnance Survey, 2008).

Die Koordinaten, welche in der Rohdatei vohanden sind, verfügen über diese Auflösung von 1 km^2. Das OSGB-Koordinatensystem würde jedoch auch eine feinere Unterteilung erlauben, in dem man die beiden Zahlengruppen in der Koordinate um weitere Ziffern ergänzt und somit eine beliebig hohe Auflösung erreichen kann. Somit widerspiegelt das oben erwähnte Beispiel mit London im Prinzip die Koordinate TQ3100081000, was einem metergenauen Standort gleichkommt. Diese Koordinate stellt somit den Quadratmeter in der linken unteren Ecke der 1 km^2 grossen Kachel dar, da der Wert der x-Achse 31'000 und der Wert der y-Achse 81'000 Meter ist innerhalb der TQ Zelle. Eine metergenaue Bestimmung eines Toponyms ist jedoch nicht möglich, da die

Koordinaten in der GEO-DB mit einer Auflösung von bloss 1 km^2 vorliegen. Auf der Website www.GEOGRAPH.org.uk bzw. in der GEO-DB wird eine 1 km^2 grosse Kachel daher über deren linke untere Ecke referenziert, wie aus der Abbildung 3.4 am Beispiel der TQ-Kachel ersichtlich ist.

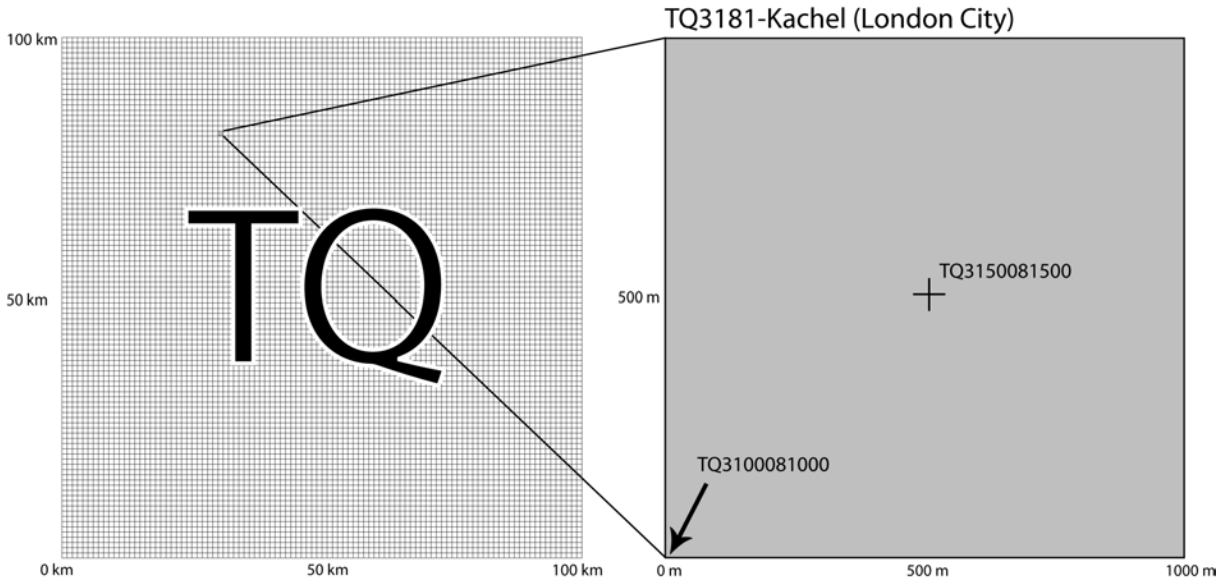


Abbildung 3.4: Zoom in die TQ Kachel. Gegebener Referenzpunkt (linke untere Ecke) vs. angenommener Referenzpunkt (Mittelpunkt).

3.3 1:50'000 Gazzetter und Einwohnertabellen von Grossbritannien

Ebenfalls von der Ordnance Survey Great Britain wurde das 50k-Gazetteer erstellt, worin alle Toponyme enthalten sind, welche auf der 1:50'000 topographischen Landeskarte von Grossbritannien engetragen sind. Diese Datenbank stellt ein Register dar, welches über 258'978 Toponyme enthält mit den entsprechenden Koordinaten. Einen Einblick in diese Datenbank gewährt die Abbildung 3.5. Da Ortsnamen teilweise unpräzise bzw. umgangssprachlich formuliert werden, können mit dem Gazetteer nicht alle Ortsbezeichnungen in der GEO-DB erfasst werden (Jones et al., 2008).

Die Einwohnertabellen von England, Schottland und Wales sind frei zugänglich auf den Webseiten der entsprechenden statistischen Ämtern publiziert. Die Einwohnertabelle von England

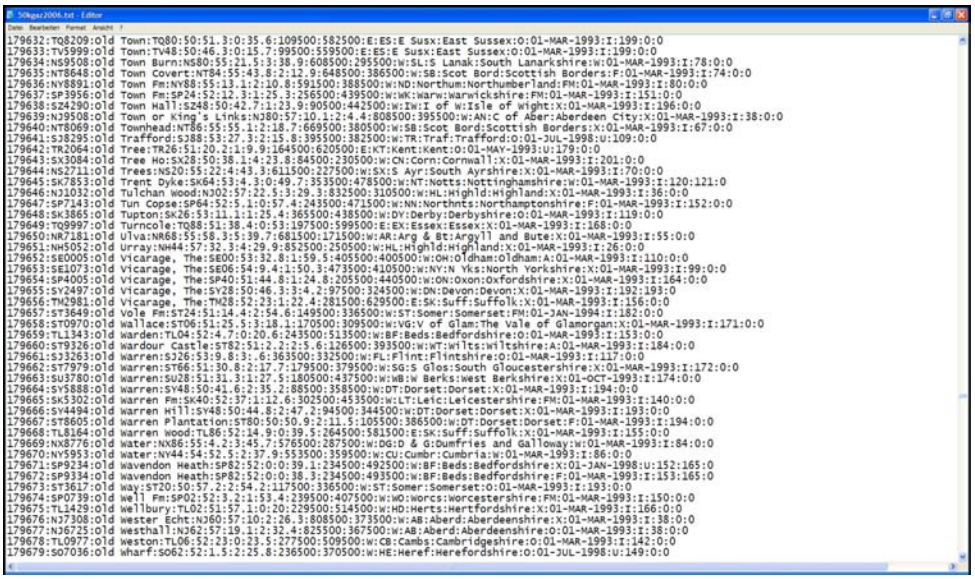


Abbildung 3.5: Ausschnitt aus dem 1:50'000 Gazetteer von Grossbritannien. ©GEOGRAPH

stellt mit einem Umfang von 2'459 Ortschaften die mächtigste Datei dar. Wales liefert 595 und Schottland 624 Einträge (www.statistics.gov.uk). Für einige Fragestellungen, bei denen die Einwohnerzahl vom Toponym mituntersucht wird, werden bloss diejenigen Beziehungen selektioniert, bei denen das Toponym auch in der Einwohnertabelle auftaucht. Dabei können sehr viele Beziehung nicht berücksichtigt werden, da lange nicht alle Toponyme eine Stadt oder ein Dorf sind.

Kapitel 4

Methodik

4.1 Manipulierung der Datenbank von GEOGRAPH

In diesem Kapitel wird erklärt, wie die Daten von GEOGRAPH durchkämmt werden, bis diese nur noch die gewünschten Informationen enthalten. Darauf wird die extrahierte Datenbank mit der Einwohnertabelle und dem Gazatteer abgeglichen und entsprechend ergänzt. Als erstes wird kurz „Perl“, die Skriptsprache, mit der die drei Datenbanken bearbeitet werden, vorgestellt.

4.2 Perl als Bearbeitungstool

Perl ist eine plattformunabhängige interpretierte Skriptsprache. Der US-amerikanische Linguist, Programmierer und Autor Larry Wall entwickelte die Sprache 1987 als Synthese hauptsächlich aus den Programmiersprachen C und awk. Ursprünglich diente die Sprache vor allem als Werkzeug, um Textdateien zu manipulieren, da unzählige reguläre Ausdrücke und Module im frei zugänglichen Comprehensive Perl Archive Network (CPAN) zur Verfügung gestellt werden. Dadurch, dass Perl gegenüber anderen Sprachen eine grösstmögliche Freiheit darstellt, hat sie auch in der Entwicklung von Webanwendung grosse Verbreitung gefunden. Perl stellt in Bezug auf Stringhandling nach wie vor ein sehr mächtiges Instrument dar und wird daher für die Verarbeitung der dieser Arbeit zu Grunde liegenden Datensätze gewählt. Die Abbildung 3.2 liefert einen Einblick in die Rohdatensätze von GEOGRAPH, deren Manipulation im folgdenen Unterkapitel behandelt wird.

4.2.1 Extraktion relevanter Informationen

Wie im Kapitel 3.3 erläutert unterliegt die Erstellung der Rohdatensätze von GEOGRAPH einer völlig anderen Motivation, als Beziehungen A *near* B auszuwerten. Die Tabelle 4.1 zeigt, wie die Datenbank über diverse Stufen so manipuliert wird, dass schlussendlich die Beziehung A *near* B vorliegt. Die einzelnen Teilschritte werden zu acht Etappen zusammengefasst. Hinter jedem Teilschritt stecken jeweils diverse Perl-Skripte. Die Teilschritte bestehen auch daraus, Fehler zu eliminieren, wie zum Beispiel unvollständig geschriebene Koordinaten des Aufnahmeortes: Teilweise haben die Fotografen bei der A-Koordinate einen Buchstaben vergessen, was später bei der Routine, welches die Distanz zwischen zwei OS-Grids berechnet, zu fehlerhaften Werten führen würde. Ein weiteres Beispiel, welches in dieser Arbeit einem Teilschritt zugeordnet wird, ist die Handhabung mit Zeilen, in denen im Kommentar und im Titel je ein Toponym vorkommt. Falls beide Toponyme identisch sind, dann wird nur eines für die spätere Auswertung verwendet, da man davon ausgehen kann, dass dabei das zweite Toponym nicht nochmals intuitiv verwendet wurde, sondern bloss eine Repetition ist. Auf die Stufen 5 bis 8 wird in den Unterkapiteln 4.2.3 und 4.2.4 näher darauf eingegangen. Am Ende dieses Verarbeitungsprozesses liegen 14'861 gültige Beziehungen in der Form wie in Abbildung 4.1 vor. Bei diesem Beispiel handelt es sich um das Toponym London, der Aufnahmestandort befindet sich in der Kachel, die 4 km nördlich von London liegt.

Aufnahmeort	räumliche Beziehung	Toponym
TQ3185	*near*	TQ3181
A	*near*	B

Abbildung 4.1: Beispiel für eine A *near* B Beziehung, wobei die Distanz zwischen A und B 4 km beträgt.

4.2.2 Ambiguität

Die Abbildung 3.5 auf Seite 21, welche einen Auszug aus dem 1:50'000 Gazetteer darstellt, zeigt die Poblematik der Ambiguität der Toponyme, wobei man von Ambiguität [lat.: ambiguus: zweifelhaft, mehrdeutig] spricht, wenn ein Wort mehrdeutig sein kann (siehe z.B. „Old Warren"). Um

Tabelle 4.1: Prozessierung der Rohdatenbanken von GEOGRAPH, dem Gazetteer und der Einwohnertabelle. (GG:GEOGRAPH, GAZ: Gazetteer, POP: Einwohertabelle, DB: Datenbank)

Stufe	Informationen	Hauptproblem	Eigenschaft	DB	Anzahl Zeilen
1	Image id, Title, Comment, Imageclass, Date of submission, Date image taken, Uploaded time, Reference index, OS A, User id, Username, Nat. eastings, Nat. northings, Nat. grlen, Viewpoint eastings, Viewpoint northings, Viewpoint grlen, View direction, Use6fig, Moderation status, Ftf, Seq no	Zu viele Informationen innerhalb einer Zeile	Vollständiger Datensatz von GG	GG	362'308
2	Image iD, Title, Comment, OS A, Imageclass, User id	Zu viel Zeilen/Datensätze	Überflüssige Angaben, wie z.B. Zeit der Bildübermittlung wurden gelöscht	GG	362'308
3	Image iD, Title, Comment, OS A, Imageclass, User id	Nicht jede Zeile enthält A *near* B, wobei B mit Grossbuchstaben beginnt	Jede Zeile enthält das Wort *near* [Regulärer Ausdruck im Perlskript: /\bnear \b/i]	GG	33'363
4	Image iD, Title, Comment, OS A, Imageclass, User id	Toponyme noch nicht identifiziert	Jede Zeile enthält den Ausdruck *near* gefolgt von einem Grossbuchstaben [Regulärer Ausdruck im Perlskript: /[A-Z]/]	GG	26'443
5	Image iD, Toponym Title, Toponym Comment, OS A, OS B, Imageclass, User id, Distanz	Ambiguität	Jede Zeile enthält ein Toponym A und B (mit Koordinate), welche im Gazetteer vorkommen. Distanz kann berechnet werden	GG, GAZ	15'635
6	Image iD, Toponym Title, Toponym Comment, OS A, OS B, Imageclass, User id, Distanz	Vereinzelt immer noch Ambiguität, welche zu Monsterdistanzen führt	Bei mehreren Toponymen B wurde immer das nächste B gewählt > fast keine Ambiguität mehr	GG, GAZ	15'635
7	Image iD, Toponym Title, Toponym Comment, OS A, OS B, Imageclass, User id, Distanz	Einwohnerzahlen fehlen	Ambiguität praktisch ausgeschlossen, da nur noch Distanzen bis 49.48 Km möglich	GG, GAZ	14'861
8	Image iD, Toponym Title, Toponym Comment, OS A, OS B, Imageclass, User id, Distanz, Einwohnerzahl	-	Toponymen, welche in der Einwohnertabelle vorkommen, wurde die Einwohnerzahl angefügt. Datensatz bereit zum Auswerten	GG, GAZ, POP	14'861

die Tragweite dieser Problematik zu veranschaulichen, will als Beispiel das Toponym „Manor Fm“ („Fm“ ist die Kurzform von „Farm“) erwähnt sein. Dieses kommt im Gazetteer 616 mal vor. 89.1% aller Toponyme im Gazetteer treten jedoch bloss nur ein einziges Mal auf, 6.4% treten zweimal auf. Bei der Ermittlung der Distanzen muss daher darauf geachtet werden, auf welches Toponym sich der Fotograf beziehen will. Die Eruierung des richtigen Toponyms stützt sich auf die Annahme, dass sich der Fotograf im Zusammenhang mit der Verwendung von *near* immer auf das nächst gelegene Toponym bezieht. Gemäss der Tabelle 4.1 wird in der Stufe 6 diese Problematik beseitigt. Ein Skript rechnet zu allen Toponymen die Distanz aus und wählt anschliessend dasjenige Toponym, welches die kürzeste Distanz zu A aufweist. Würde man die Problematik der Ambiguität ignorieren, hätte dies gravierende Auswirkungen auf die Resultate, da so riesige Distanzen von bis über 1000 Kilometer in die Auswertungen fliessen würden, was die Resultate stark verfälschen würde.

Nach einer ersten groben Elimination der Ambiguität tritt dieses Problem vereinzelt immer noch auf. Selten tritt der Fall ein, dass sich der Fotograf auf ein Objekt bezieht, welches nicht im Gazetteer registriert ist, jedoch mit einem Eintrag identisch bzw. ambig ist. Der Teilnehmer könnte sich zum Beispiel auf eine Person beziehen, deren Name gleich lautet wie ein Toponym im Gazetteer. Um diese einzelnen Ausreisser zu beseitigen, muss ein Grenzwert bezüglich Distanz definiert werden. Dieser wird dort angesetzt, wo der kumulierte Prozentwert aller Distanzen bei exakt 95% liegt, was einem 0.05-Quantil entspricht. Die restlichen Distanzen, welche noch weniger als 5% ausmachen, werden aus der Datenbank enfernt. Dieser Grenzwert scheint eine vernünftige Lösung für diese Problematik zu sein, wenn man bedenkt, dass dies einem Spektrum von 0 bis 49.48 Kilometer entspricht. Dabei wird angenommen, dass es realistisch ist, Distanzen bis rund 50 Kilometer im Zusammenhang mit *near* zu verwenden und gleichzeitig, dass ab dieser Distanz *near* nur noch äusserst selten angewendet wird im Zusammenhang mit Bildbeschreibungen. Die später vorgestellten Untersuchungen in dieser Grössenordnung, spielen sich innerhalb eines 71x71 Km Rasters ab, wobei immer vom zentralsten Quadrat aus zu seinen umliegenden Zellen gemessen wird. Dabei stellt die Distanz vom Mittelpunkt zu den Eckpunkten die maximale Distanz von 49.48 Km dar. Nach dieser Extraktion der unerwünscht grossen Distanzen ist die Datenbank nun bereinigt und soweit aufbereitet, dass sie mit der Einwohnertabelle und dem Gazetteer abgeglichen und dementsprechend erweitert bzw. bereichert werden kann. Das folgende Kapitel geht auf diese eben erwähnten Schritte näher ein.

4.2.3 Abgleichen von Datensätzen

Wie in der Tabelle 4.1 ersichtlich ist, wird ab Stufe 5 das im Kapitel 3.3 erläuterte Gazetteer von Grossbritannien miteinbezogen. Dabei sorgt ein Perlskript dafür, dass auf die Datei von GEO-

GRAPH und auf das Gazetteer zugegriffen werden kann (siehe Beispielskript im Anhang). Das Skript ist so geschrieben, dass jeweils nur noch diejenigen Zeilen in einer neuen Datei ausgegeben werden, wenn das Toponym von der GEO-DB im Gazatteer vorkommt. Dieser Prozess führt zu einer Reduktion des Datensatzes auf 15'635 Zeilen. In der Stufe 8 werden die Datensätze mit der Einwohnertabelle abgeglichen. Falls ein Toponym in beiden Daten auftritt, dann wird die GEO-DB mit der entsprechenden Einwohnerzahl erweitert, kommt das Toponym nicht in der Einwohnertabelle vor, dann besteht diese Zeile ohne einen Eintrag bei der Einwohnerzahl weiterhin. In 3'327 von 15'635 Fällen konnte die Datenbank mit dieser Information bereichert werden. Der nächste Schritt besteht darin, die Distanzen zwischen der Koordinate des Aufnahmeortes und dem Zielobjekt zu berechnen.

4.2.4 Distanzberechnung

Um die Distanzberechnung nachvollziehen zu können, wird das Verständnis der Systematik des bereits erklärten Koordinatensystems von Grossbritannien im Kapitel 3.3 vorausgesetzt. Wie in der Tabelle 4.1 zu sehen ist, wird in der Verarbeitungsphase der Datenbank von GEOGRAPH während der Stufe 5 die Distanz zwischen der gegebenen OS Kachel vom Standort A und der OS Kachel des Toponyms B berechnet. Grundsätzlich konvertiert das Perlskript die zwei Buchstaben der OS Koordinate in einen quantitativen Wert und setzt diesen mit den zwei Ziffernblöcken so zusammen, dass die daraus entstehende neue Koordinate mathematisch verwendet werden kann. Diese Umwandlung wird mit der OS Koordinate des Aufnahmeortes und des Toponyms durchgeführt, somit kann für jede Beziehung, über die einfache Funktion des Pythagoras, die euklidische Distanz berechnet werden. Einen Einblick in das Skript, welches diese Distanzberechnung bewerkstelligt, wird im Anhang aufgeführt. Die Grafik 4.2 zeigt ein Beispiel eines schematischen Ablaufs der Umwandlung der OS Kacheln in rechenbare Koordinaten.

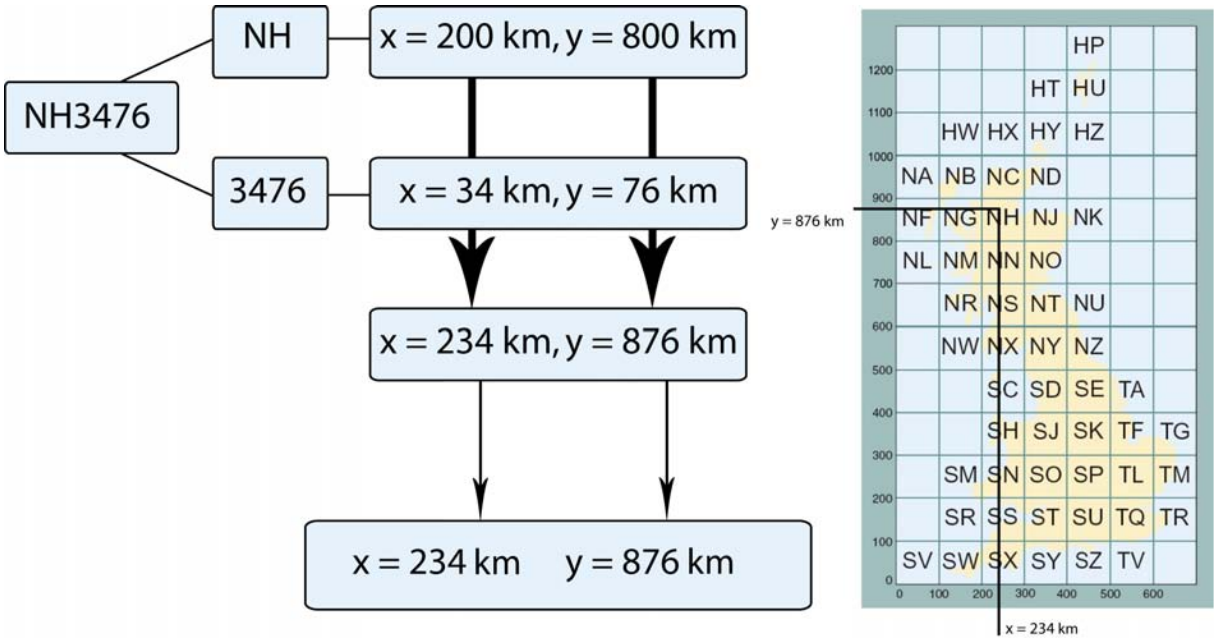


Abbildung 4.2: Schematische Darstellung der Konvertierung der OS Kacheln in rechenbare Koordinaten (Karte: Ordnance Survey, 2008)

4.3 Metriken für Rasterdaten

Die Rastermetrik spielt im Zusammenhang mit dieser Arbeit eine zentrale Rolle, da für die Auswertung der Distanzen ein (vorgegebenes) Raster verwendet wird. Da das Koordinatensystem der Ordnance Survey einem Raster unterliegt, stützen sich auch die statistischen Auswertungen der Koordinaten auf einer rasterbasierten Metrik. Eine Metrik bedeutet eine Abstandsfunktion $D(a, b)$ zwischen zwei Punkten a und b. Um die Distanz zwischen den Punkten $a = (a_x, a_y)$ und $b = (b_x, b_y)$ in einem Raster zu messen, kommen grundsätzlich drei Methoden in Frage (Bill, 1999):

- Euklidische Metrik ($D_e(a, b)$)
- City-Block Metrik ($D_c(a, b)$)
- Schachbrett Metrik ($D_s(a, b)$)

Die erste Methode basiert auf der euklidischen Distanzmessung, welche die direkte Verbindung zwischen zwei Punkten misst. Eine andere Möglichkeit, eine Distanz zu messen, ist die City-Block- (oder Manhatten-) Metrik. Dabei wird der horizontale und vertikale Abstand zwischen den Punkten gemessen und summiert. Die dritte Methode beruht auf der Schachbrett-Metrik,

bei welcher der Abstand über das Maximum des horizontalen und vertikalen Abstandes definiert wird. Diese verschiedenen Distanzfunktionen werden wie folgt dargestellt (Tang, 1991; Lichtner, 1981):

$$D_e(a,b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$$

$$D_c(a,b) = |(a_x - b_x)| + |(a_y - b_y)|$$

$$D_s(a,b) = max(|(a_x - b_x)|, |(a_y - b_y)|)$$

Die Abbildung 4.3 veranschaulicht die Geometrie der drei verschiedenen Messarten und die daraus resultierenden Distanzen in einem quadratischen 5x5 Raster. Die Distanzen der euklidischen, der City-Block und der Schachbrett Metrik werden jeweils vom zentralsten Quadrat (auch Zelle genannt) aus zu seinen 24 umliegenden Zellen berechnet.

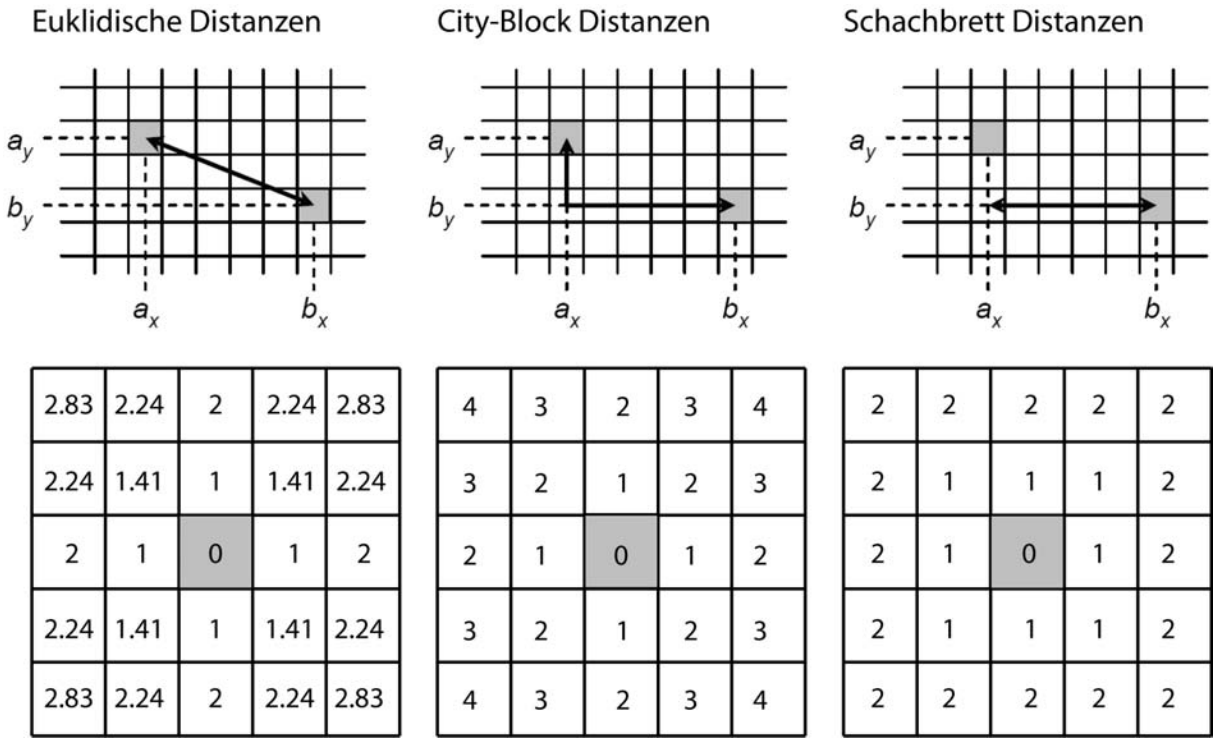


Abbildung 4.3: Vergleich der drei verschiedenen Metriken im Raster (Bill, 1999)

Aus den vorherigen Abbildungen wird ersichtlich, dass eine Zelle bei der City-Block Distanz bloss das Quadrat links, rechts, ober- und unterhalb als seine Nachbarn betrachtet. Diese Beziehung wird als eine 4-er Nachbarschaft bezeichnet, da jedes Quadrat vier Nachbarn hat. Bei

der Schachbrett-Messung werden jedoch die Quadrate, welche diagonal angrenzen, ebenfalls als Nachbarn definiert. Dabei wird allen umliegenden acht Rastern dieselbe Distanz zugeordnet, wodurch man von einer 8-er Nachbarschaft spricht (Lichtner, 1981). Beim letzteren Nachbarschaftstyp kann man erkennen, dass sich in Falle eines 5x5 Rasters bloss 4 Klassen bilden, wobei sich beim 4-er Nachbarschaftstyp (mit der euklidischen Distanzmessung) sechs Abstufungen ergeben und dieser somit eine feinere Granularität besitzt. In der Tabelle 4.2 werden anhand eines Beispiels zweier Punkte die unterschiedlichen Distanzen auf Grund der verschiedenen Metriken aufgezeigt. Das Beispiel wird mit den Punkten $a(5, 5)$ und $b(25, 17)$ durchgeführt.

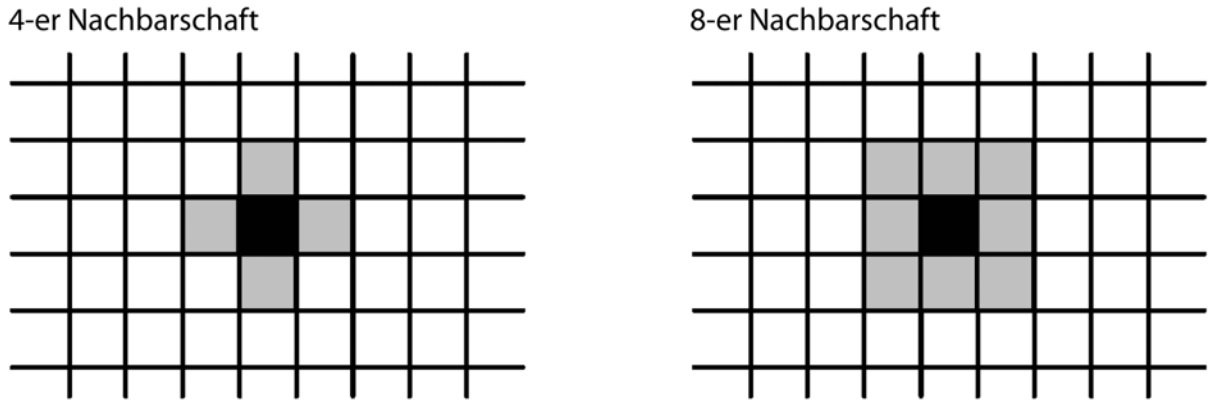


Abbildung 4.4: Nachbarschaftstypen in einem quadratischen Raster (Lang, 2005)

Tabelle 4.2: Abstandsberechnungen mit verschiedenen Distanzfunktionen (Tang, 1991)

Distanzfunktion	Abstand	Differenz
Euklid	23.324	-
City-Block	32.000	8.676
Schachbrett	20.000	-3.24

Die Koordinaten im Datensatz von GEOGRAPH basieren ebenfalls auf einem quadratischen Raster mit je 1 km^2 grossen Zellen. Da die euklidische Distanzfunktion gemäss Tabelle 4.2 die präziseste Messart darstellt und eine feinere Abstufung erzielt, ist die Wahl dieser Metrik in dieser Diplomarbeit eine plausible Entscheidung. Da die Auflösung der OS Kacheln mit 1 km^2 grundsätzlich nicht sehr hoch ist, wird somit die bestmögliche Auflösung bewahrt.

4.3.1 Metrik in einem 5x5 km Raster

Gemäss den vorhergehenden Formeln der verschiedenen Metriken ist die kleinste Distanz, welche bei einer Berechnung zweier Koordinaten auftreten kann 0 km. In dem Fall, wenn sich der

Fotograf in derselben Kachel befindet, wie das Toponym, welches er als naheliegend bezeichnet. 1, 1.41, 2, 2.24 und 2.83 km bilden die weiteren Distanzen, welche unter anderem in einem 5x5 km Raster untersucht werden. Diese fünf Distanzen entstehen, wenn man vom zentralsten Feld aus zu seinen 24 umliegenden Feldern die euklidische Distanzmessung anwendet (siehe Abb. 4.3, S. 28).

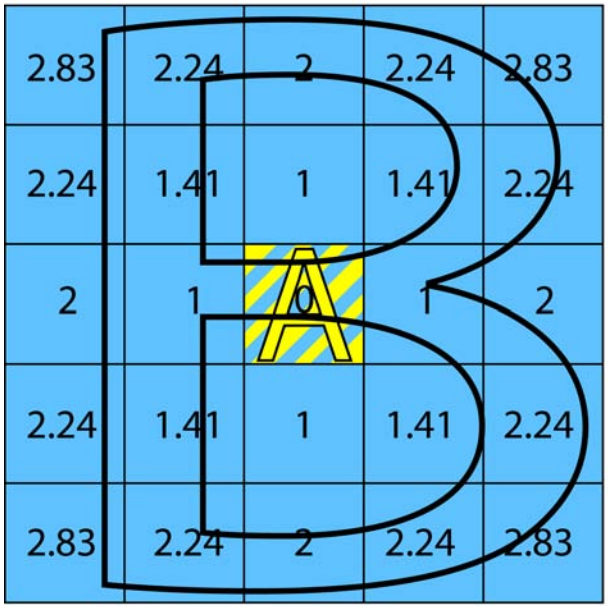


Abbildung 4.5: A befindet sich im zentralsten Quadrat, B kann sich in allen 25 Zellen befinden, Distanzen in km.

Von allen untersuchten Beziehungen der Datensätze der GEO-DB spielen sich 85.3% innerhalb des Bereiches von 0 bis und mit 2.83 km ab, wobei 2.83 km lediglich noch bei 1.6% aller Fälle auftaucht.

Die Distanzen werde immer von der linken unteren Ecke eines Quadrates zur linken unteren Ecke des anderen Quadrates berechnet. Die Fixierung dieser beiden Punkte an der linken unteren Ecke ist auf Grund der Systematik des Koordinatensystems in Grossbritannien so definiert (vgl. Abb. 3.4, S. 20). Mathematisch ergibt diese Berechnungsmethode die selben Resultate, wie wenn man als Referenzpunkt den Mittelpunkt annehmen und immer vom Mittelpunkt des einen Quadrates zum Mittelpunkt des anderen Quadrates messen würde und sich somit auf die so genannte Methode des Zentroid-Abstandes stützen würde.

Es wird angenommen, dass sich die Fotografen tendenziell eher in der Mitte als in der linken unteren Ecke einer quadratischen Fläche befinden. Diese Mittelpunktannahme stellt eine Vereinfachung der Distanzberechnung dar. Im Prinzip müsste jedoch bei jeder Distanzberechnung der Range bzw. der Bereich der potentiellen Distanzen berücksichtigen. Die Distanz 0 km bedeutet, dass A und B im selben Quadrat liegen. Theoretisch könnten A und B im extremsten Fall jedoch auch 1.41 km weit auseinander liegen, falls sich A und B in der diagonal gegenüberliegenden Ecke befinden. Die Wahrscheinlichkeit, dass dies so wäre, ist jedoch sehr klein. Jede Distanz zwischen 0 und 1.41 km hat eine gewisse Eintrittswahrscheinlichkeit, je nachdem, wie oft eine Distanz in einem 1 km^2-Quadrat platziert werden kann. Daher müsste man für alle möglichen Distanzen einen Gewichtungsindex erstellen. Dies wäre jedoch sehr kompliziert und würde vermutlich keinen grossen Unterschied ausmachen im Gegensatz zu den Resultaten, die auf der Mittelpunktannahme basieren. Denn die Strecke von Mittelpunkt zu Mittelpunkt kann in einer Zelle am meisten abgetragen werden; im Gegensatz zu längeren oder kürzeren Strecken, bei denen die Wahrscheinlichkeit mit zunehmender Abweichung von der Mittelpunktdistanz steigt, dass ein Punkt nicht mehr im gewünschten Quadrat platzierbar ist.

Kapitel 5

Resultate

Die folgenden Resultate zeigen auf, welche Distanzen mit der Verwendung von *near* im Zusammenhang mit Bildbeschreibungen in verschiedenen Regionen in Grossbritannien entstanden sind. Es werden einzelne Regionen miteinander Verglichen, wobei auf unterschiedliche Distanzspektren fokussiert wird. Zudem wird untersucht, ob eine Korrelation mit der verwendeten Distanz mit *near* und der Einwohnerzahl vom Toponym B zu erkennen ist. Der Schluss dieses Kapitel wird durch Vergleiche von zufälligen Verteilungen der Toponyme vom Gazetteer und denjenigen Verteilungen von Toponymen aus der GEO-DB gebildet.

5.1 Statistische Tests

5.1.1 Kolmogorov-Smirnov-Test

Viele statistische Tests setzen eine mindestens annähernde Normalverteilung der Proben voraus. Daher werden die Stichproben in dieser Arbeit, welche aus Distanzen bestehen, mit einem Kolmogorov-Smirnov-Test auf Normalverteilung überprüft. Dieser nichtparametrische Test gilt als sehr stabil und wurde ursprünglich für metrische Merkmale entwickelt, weshalb er sich auch für die ebenfalls metrisch skalierten Stichproben aus der GEO-DB eignet. Nichtparametrische (oder verteilungsfreie) Tests verlangen keine Verteilungsvoraussetzungen und sind somit mächtiger als andere Methoden (Ebdon, 1977).

Der Kolmogorov-Smirnov-Test generiert mit den vorliegenden Proben bezüglich Normalverteilung eine Signifikanz von kleiner als 0.001. Dieser Wert ist in diesem Zusammenhang nicht akzeptabel, weshalb man keine Normalverteilung annimmt. Das Q-Q Diagramm in Abbildung 5.1 veranschaulicht dieses Ergebnis: Die gestrichelte Gerade repräsentiert eine perfekte Normalverteilung. Würden die Werte eng um diese Gerade streuen, könnte man von einer Normalver-

teilung sprechen.

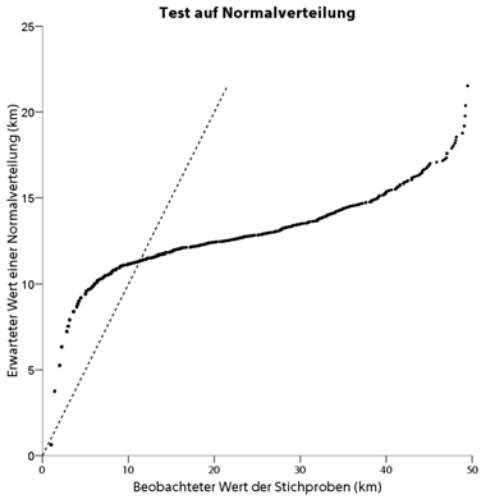


Abbildung 5.1: Q-Q-Diagramm der Distanzen von 0 bis 49.48 km.

5.1.2 Mann-Whitney-U-Test

Es wird eine zufällig generierte Häufigkeitsverteilung und die Häufigkeitsverteilung, welche auf den Datensätzen von GEOGRAPH beruhen, auf ihre Übereinstimmung überprüft. Stimmen beide überein, würde dies aussagen, dass die Verteilungen auf einem bestimmten Signifikanzniveau (Irrtumswahrscheinlichkeit) aus derselben Grundgesamtheit stammen. Daraus könnte man ableiten, dass die Verteilung der Stichproben aus der GEO-DB ebenfalls eine zufällige Verteilung und somit uninteressant für weitere Untersuchungen wäre. Es wird jedoch erwartet, dass die verschiedenen Stichproben resp. Distanzen zwischen A und B nicht zufällig generiert wurden und daher weitere Untersuchungen interessant bleiben.

Um diese Unabhängigkeit zu überprüfen eignet sich der nichtparametrische Mann-Whitney-U-Test. Im Gegensatz zu anderen statistischen Tests, wie zum Beispiel dem Student's t-Test, setzt er keine Normalverteilung der Proben voraus. Die Nullhypothese besagt, dass zwischen den beiden Stichproben mit n_x und n_y Werten kein Unterschied existiert sie und somit aus der gleichen Grundgesamtheit stammen. Jeder Wert der zufälligen Verteilung wird mit jedem Wert der GEO-DB verglichen. Bei dem Distanzspektrum von 0 bis 2.83 km ergibt dies insgesamt 12'671 x 12'671 Vergleiche. Diese Distanzen spielen sich in einem 5x5 km Raster ab, wobei sechs verschiedene Distanzen mit unterschiedlichem Flächenanteil entstehen. Der Flächenanteil pro Distanz

widerspiegelt die Wahrscheinlichkeit bei einer zufälligen Verteilung. Die Distanz 0 km tritt bloss in einem Feld auf, das heisst, diese Distanz tritt mit einer Wahrscheinlichkeit von 1/25 ein. 1, 1.41, 2 und 2.83 km kommen je viermal vor; dies entspricht einer Wahrscheinlichkeit von 4/25, die Distanz 2.24 km tritt acht mal auf. In der Tabelle 5.1 sind die Werte einer zufälligen Verteilung in einem 5x5 km Raster aufgelistet. In der Abbildung 5.2 wird diese zufällige Verteilung mit der Verteilung gemäss GEO-DB in einem Quervergleich abgebildet.

Tabelle 5.1: Werte einer zufällige Verteilung der Stichproben von 0 bis 2.83 km. P(x) bedeutet die Wahrscheinlichkeit, dass die Distanz x eintritt, $P(x)$ x 12′671 ist die Anzahl Stichproben pro Distanz.

Distanz x in km	P(x)	$P(x)$ x 12′671
0	1/25	506.84
1	4/25	2027.36
1.41	4/25	2027.36
2	4/25	2027.36
2.24	8/25	4054.72
2.83	4/25	2027.36
		$\sum$ 12’671

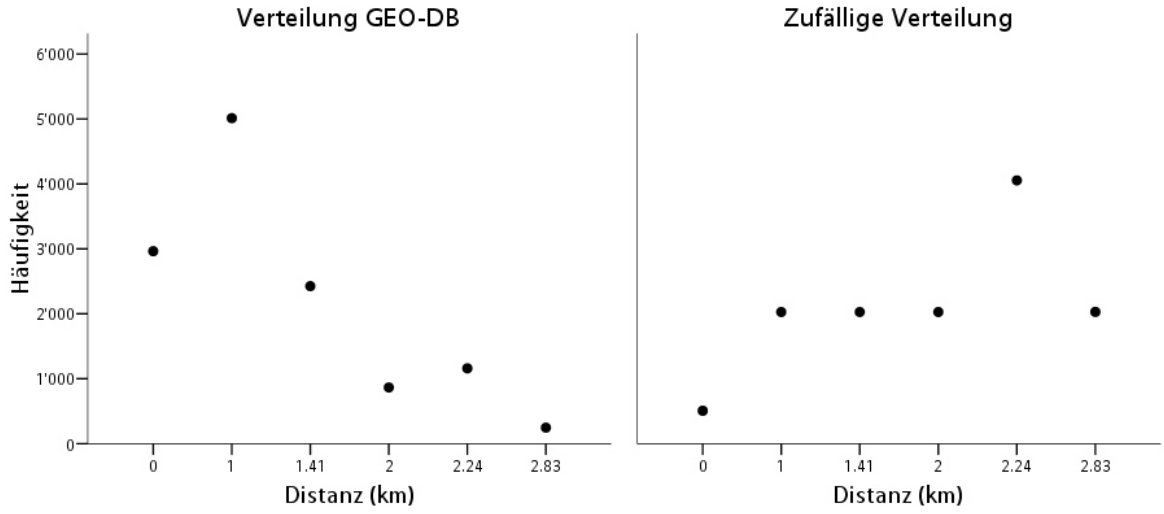


Abbildung 5.2: Die Häufigkeitsverteilung der GEO-DB für Distanzen bis 2.83 km (links) und eine zufällige Häufigkeitsverteilung.

Um den Mann-Whitney-U-Test nun durchzuführen gilt es, den U-Wert (U_{emp}) zu berechnen: r_x und r_y stellen die Rangzahlen der geordneten Distanzen der beiden Verteilungen dar.

Die Rangzahlen werden getrennt in zwei Spalten zur so genannten Rangsumme aufsummiert ($\sum r_x, \sum r_y$). Falls mehr als zwei Zahlenwerte identisch sind, werden in beide Rangspalten jeweils die arithmetischen Mittel eingetragen. Der Mann-Whitney-U-Test greift somit nicht direkt auf die Werte der Stichproben zurück, er stützt sich bloss auf deren Rangordnung. Die Formel für U_{emp} lautet folgendermassen:

$$U_{emp} = Min\left[n_x n_y + \frac{n_x(n_x+1)}{2} - \sum r_x, n_x n_y + \frac{n_y(n_y+1)}{2} - \sum r_y\right]$$

Löst man die Formel mit den insgesamt 2 x 12'671 Stichproben auf, gibt sie einen U-Wert von nahezu Null zurück. Diesen Wert muss man mit dem kritischen U-Wert (U_{crit}) vergleichen. Da $U_{crit} > U_{emp}$ ist, kann die Nullhypothese nun verworfen und angenommen werden, dass die beiden Stichproben mit einem Signifikanzniveau von 0.05 stark signifikant unterschiedlich sind und somit nicht aus derselben Grundgesamtheit stammen. Dies bedeutet, dass die Verteilung der Stichproben aus der GEO-DB nicht zufällig generiert wurde. Die Häufigkeitsverteilungen der drei weiteren untersuchten Distanzspektren werden ebenfalls mit einem Mann-Whitney-U-Test auf ihre Unabhängigkeit gegenüber einer zufälligen Verteilung getestet. Die Resultate sind in der Tabelle 5.2 einzusehen. Für alle vier untersuchten Distanzspektren kann die Nullhypothese verworfen und angenommen werden, dass keine zufällig Verteilung in den Daten von GEOGRAPH existiert.

Tabelle 5.2: Mann-Whitney-U-Test mit verschiedenen Distanzspektren. p: Irrtumswahrscheinlichkeit bei $\alpha = 0.05$.

Distanz in km	n_x, n_y	p
0-49.48	14'861	$p < 0.05$
0-11.31	14'386	$p < 0.05$
3-49.48	2'190	$p < 0.05$

5.2 Die Bedeutung von *near* in ganz Grossbritannien

Um einen visuellen Eindruck zu erhalten, wie die Stichproben aus dem Datenbestand von GEOGRAPH in Grossbritannien verteilt sind, werden die OSGB-Koordinaten mittels einem Perlskript in rechenbare Koordinaten umgewandelt und anschliessend in das GI-Programm „ArcMap“ geladen, um eine Dichtefunktion der Aufnahmeorte zu generieren. Die linke Grafik in der Abbildung 5.3 präsentiert alle 362'308 Aufnahmeorte, die in der GEO-DB vorkommen (auch jene, die keine „*A near B*“-Beziehung enthalten). In der rechten Grafik sind nur noch jene 14'861 Aufnahmeorte von Datensätzen abgebildet, welche die gültige Beziehung *A near B* enthalten. Die Dichtefunktionen werden wie folgt erzeugt:

Mittels *Point Density*-Verfahren wird eine Dichte berechnet, bei der bloss Punkte berücksichtigt werden, die in einer bestimmte Nachbarschaft auftreten (nearest neighbor for discret data). Folgende Parameter werden benutzt:

- Output cell size: 1
- Neighborhood: Circle
- Radius: 20 km
- Unit: Map
- Area units: Square kilometers

Nach der Erstellung der Dichtefunktion wird der Output mittels der Funktion *Extract by mask* mit einem Polygon in der Form von Grossbritannien maskiert. Der Output besteht aus *floating points*, was die approximative Darstellung der reellen Punktmenge bedeutet. Die Klassierung beruht auf einem 5%-Quantil, d.h. es treten in allen 20 Klassen jeweils 5% aller Stichproben auf.

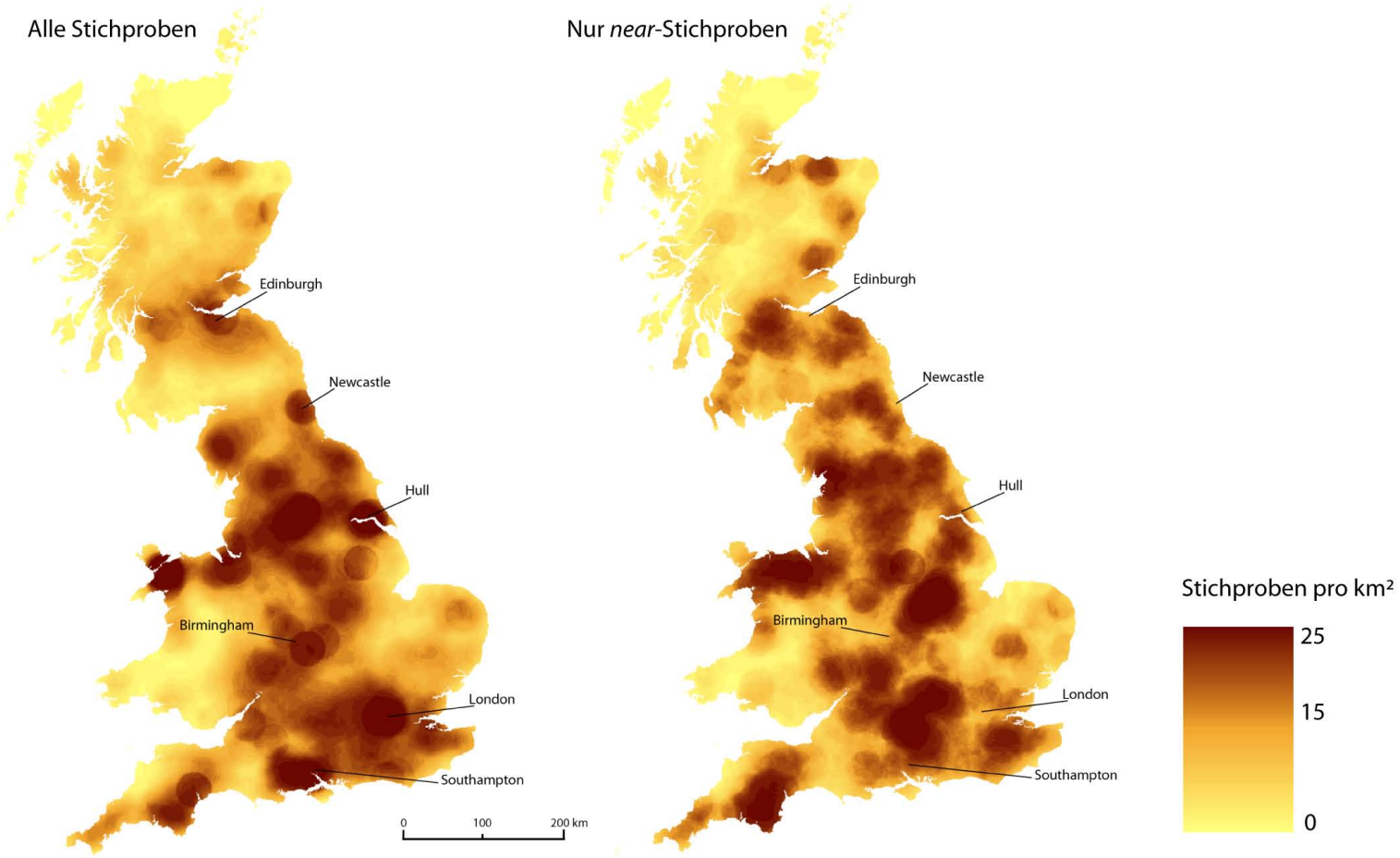


Abbildung 5.3: Dichtefunktion aller 362'308 Aufnahmeorte von GEOGRAPH (links) und nur jene 14'861, die im Zusammenhang mit *near* verwendet werden.

In den folgenden Unterkapiteln werden zuerst die grösseren Distanzen, die in ganz Grossbritannien auftauchen, untersucht, danach die kleineren. Anschliessend werden einzelne Regionen näher betrachtet, wobei das Distanzspektrum je nach Grösse des Ausschnittes und der Anzahl Stichproben variiert. Später wird untersucht, ob Korrelationen sichtbar sind, zum Beispiel zwischen der Distanz und der Einwohnerzahl vom Toponym B. Zum Schluss dieses Kapitels wird getestet, wie sich die Verteilung von GEOGRAPH von einer Verteilung unterscheidet, welche dadurch entsteht, in dem man jeweils der gegebenen Koordinate A eine zufällige Koordinate B aus dem Gazetteer zuordnet.

5.2.1 Distanzen von 0 bis 49.48 km

Die meisten Distanzen spielen sich in einem sehr engen Bereich ab: 53.7% der 14'861 Distanzen sind 0 oder 1 km, wobei letztere Distanz 33.7% repräsentieren. Ab 1.41 km nimmt die Häufigkeit pro Distanz stark ab, so dass die Häufigkeitsverteilung in der Abbildung 5.4 logarithmisiert werden muss, damit man auch bei grösseren Distanzen differenzierte Werte auf der x-Achse wahrnehmen kann. Bei einer linearen Verteilung wäre die Funktion ab etwa 10 km nicht mehr von der x-Achse zu unterscheiden. Die komplette Häufigkeitstabelle aller existierenden Distanzen ist im Anhang zu finden.

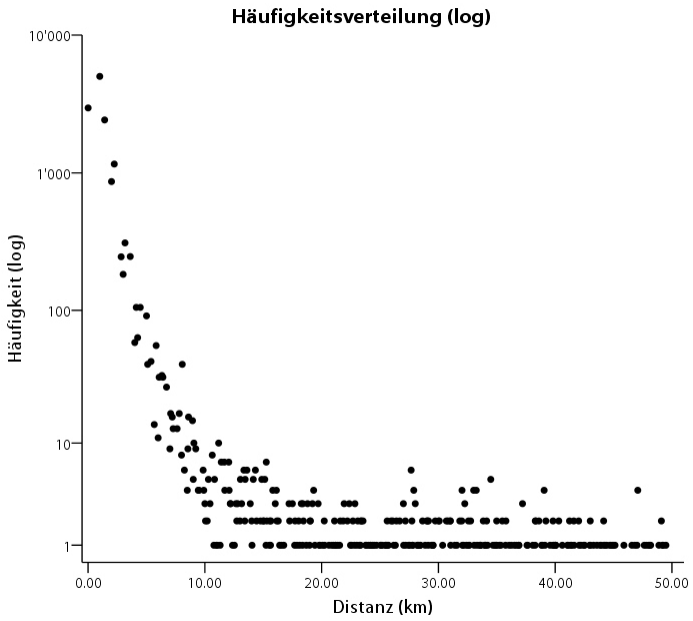


Abbildung 5.4: Häufigkeitsverteilung aller gültigen Stichproben von Grossbritannien.

In der Tabelle 5.3 sind weitere statistische Kenngrössen ersichtlich, die sich im Zusammenhang mit der Verteilung der Distanzen von 0 bis 49.48 km ergeben.

Tabelle 5.3: Statistische Kenngrössen der Distanzen von 0 bis 49.48 km.

Anzahl Proben	14'861
Mittelwert	2.28 km
Median	1 km
Standardabweichung	4.89 km

Die logarithmisierte Häufigkeitsverteilung in Abbildung 5.4 zeigt einen negativ exponentiellen Verlauf. Die Frage stellt sich, ob dieser Verlauf so ist, da das Intervall der Unterteilung der Distanzen kleiner wird, je höher die Distanz ist. Als Veranschaulichung dieses Problems betrachte man das Spektrum zwischen 0 und 2.83 km - in diesen 25 Quadraten tauchen sechs verschiedene Distanzen auf. Wird der Fokus nun beispielsweise auf einen gleich grossen Ausschnitt (2.83 km) gerichtet, jedoch zwischen 46.65 und 49.48 km, dann sind viel mehr Quadrate von diesem Spektrum betroffen und somit entstehen auch viel mehr mögliche Distanzen. Damit ist die Wahrscheinlichkeit kleiner, dass viele Proben auf ein und dieselbe Distanz zutreffen, als im Bereich von 0 bis 2.83 km.

Es wird nun untersucht, ob sich eine Häufigkeitsverteilung unter Berücksichtigung dieses Umstandes wesentlich von der Verteilung in Abbildung 5.4 unterscheidet oder ob sich die beiden Verteilungen ähnlich sind. Um den Effekt zu neutralisieren, dass Ausschnitte gleicher Grösse je nach Entfernung vom Aufnahmestandort mehr oder weniger Ergebnisse enthalten („Selektions-Inhomogenität"), wird eine Verteilung generiert, bei welcher die Distanzen in 5 km-Stufen zusammengefasst bzw. klassiert werden. Somit erhält jedes (gleich grosse) Distanzspektrum einen Häufigkeitswert, welcher unterschiedlich viele Distanzen in einer Klasse vereint. Die Abbildung 5.5 stellt diesen Sachverhalt in einer logarithmisierten Funktion dar.

5.2.2 Distanzen von 0 bis 11.31 km

Innerhalb der Distanz von 0 bis 11.31 km spielen sich 96.8% aller gültigen Beziehungen ab. Die Besonderheit an diesem Spektrum liegt darin, dass die restlichen 3.2% der Distanzen zwischen 11.31 und 49.48 km auch als Ausreisser verstanden werden können. Es versteht sich, dass bereits ein Ausreisser gewisse stichprobenbasierte Resultate stark beeinflussen kann. Im folgenden Kapitel werden regionale Durchschnittswerte ermittelt und analysiert, welche zum Teil bloss auf 10 Stichproben basieren. Die Ergebnisse dieser Stichproben würden auf solche Ausreisser höchst sensibel reagieren. Daher werden für diese Untersuche genau diejenigen Proben mit maximalen Distanzen bis 11.31 km verwendet. Wie die Verteilung von Distanzen bis 11.31 km ausschaut,

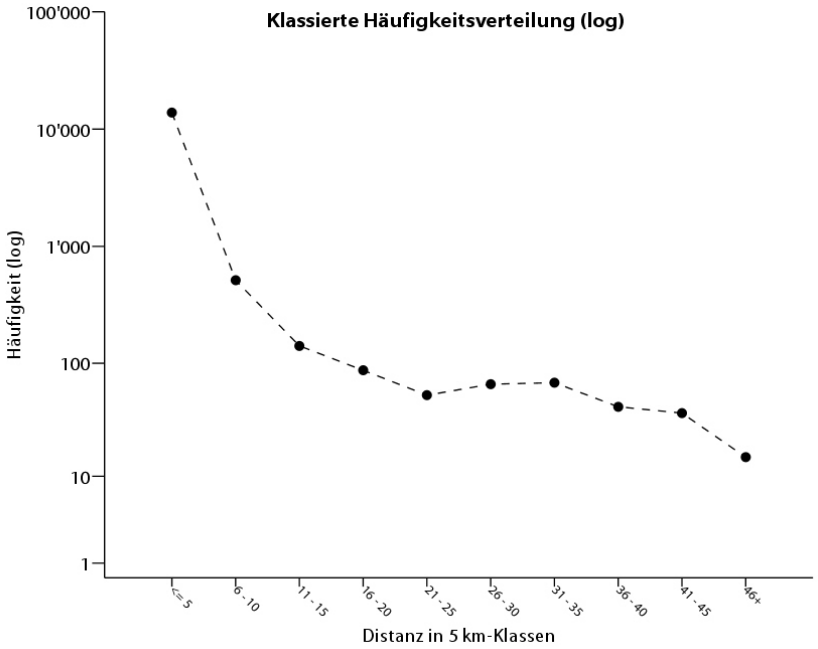


Abbildung 5.5: Logarithmisierte Häufigkeitsverteilung für Distanzen von 0 bis 49.48 km, klassiert in 5 km-Klassen.

wird in der Abbildung 5.6 aufgezeigt. Auch dieses Diagramm muss aus demselben Grund, wie bei der oben erwähnten Verteilung, logarithmisiert werden. Um die Notwendigkeit einer Logarithmisierung zu veranschaulichen, wird hier zudem auch noch die lineare Funktion abgebildet. Ab etwa 4 km kann man den x-Werten bei der nicht logarithmisierten Darstellung keine differenzierten y-Werte mehr zuordnen.

Der Mittelwert, der Median und die Standardabweichung dieser Verteilung sind in der Tabelle 5.4 zu sehen. Man erkennt im Vergleich zur Tabelle 5.3 (die Distanzen von 0 bis 49.48 km berücksichtigt), dass der Mittelwert nur relativ geringfügig ändert, obwohl die untersuchte Distanz wesentlich kleiner ist. Dagegen unterscheidet sich, wie zu erwarten ist, die Standardabweichung ganz klar gegenüber derjenigen von der Tabelle 5.3.

Tabelle 5.4: Statistische Kenngrössen der Distanzen von 0 bis 11.31 km.

Anzahl Proben	14'386
Mittelwert	1.51 km
Median	1 km
Standardabweichung	1.55 km

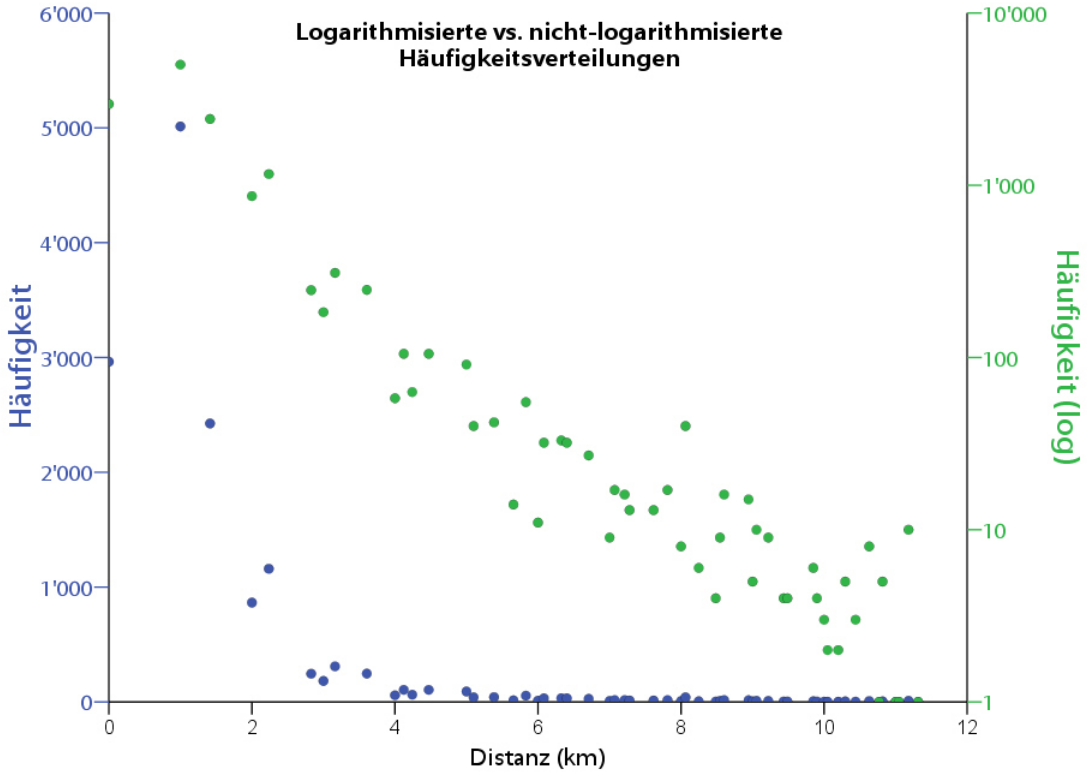


Abbildung 5.6: Häufigkeitsverteilung der Proben mit Distanzen von 0 bis 11.31 km. Blau: nicht logarithmisierte Häufigkeit, grün: logarithmisierte Häufigkeit.

5.2.3 Distanzen von 0 bis 2.83 km

Von den insgesamt 14'861 Stichproben treten 85.3% zwischen 0 und 2.83 km auf. Die Abbildung 5.7 zeigt die entsprechende Häufigkeitsverteilung. Deutlich zu erkennen ist, dass am meisten Proben in den direkt angrenzenden Kacheln des OSGB-Koordinatensystems auftreten, was genau die Distanz von 1 km bedeutet.

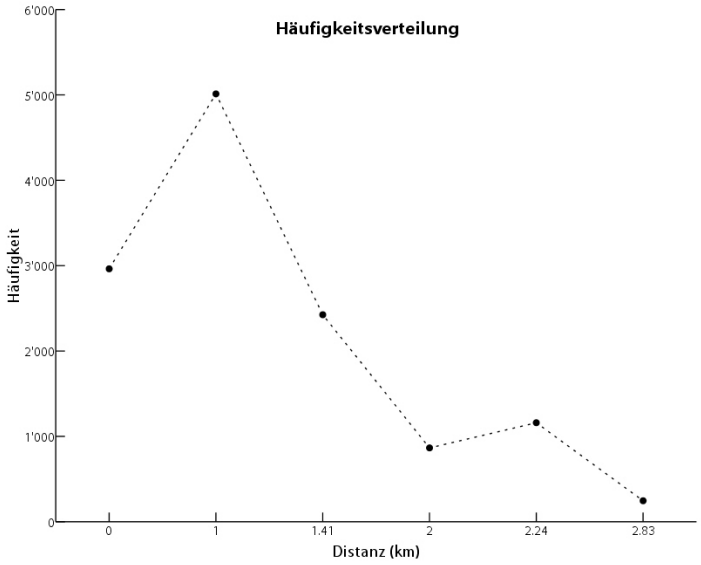


Abbildung 5.7: Häufigkeitsverteilung der Distanzen von 0 bis 2.83 km.

Bei einer deskriptiven Datenanalyse dieser Verteilung ergeben sich die statistischen Kenngrössen, welche in der Tabelle 5.5 aufgeführt sind. Stellt man einen Vergleich mit den Werten in der Tabelle 5.4 der Verteilung vom Abschnitt 5.2.2 an, kann festgestellt werden, dass der Median, wie bei allen Distanzen, unverändert bei 1 km bleibt, der Mittelwert verringert sich dagegen um rund 0.5 km. Ebenfalls kleiner wird, auf Grund des schmaleren Spektrums, die Standardabweichung.

Tabelle 5.5: Statistische Kenngrössen der Distanzen von 0 bis 2.83 km.

Anzahl Proben	12'671
Mittelwert	1.06 km
Median	1 km
Standardabweichung	0.73 km

5.2.4 Distanzen von 3 bis 49.48 km

Nachfolgend wird aufgezeigt, wie sich die Häufigkeiten aller Distanzen verhalten, die von 3 bis und mit 49.48 km auftreten. Dieses Spektrum ist insofern interessant, da sich damit das Verhalten der „Randwerte“ beobachten lässt. Es treten lediglich 14.7% (bzw. 2'192) aller Stichproben in diesem Segment auf, wobei der betrachtete Ausschnitt mit 46.65 km 94.3% der gesamten Distanz einnimmt. Im Kapitel 4.2.2 wurde erwähnt, dass die Beziehungen, welche Distanzen bis 49.48 km enthalten, innerhalb eines 71x71 km Rasters untersucht werden. Diese Stichproben weisen von 660 möglichen Distanzen, welche in einem 71x71 km Raster mit 5'016 Zellen untersucht werden, deren 314 verschiedene Ausprägungen auf. Dabei blendet man die innersten 25 Zellen, die einem

5x5 km Raster entsprechen, aus. Die Abbildung 5.8 zeigt die Häufigkeiten, welche zwischen einer Distanz von 3 bis und mit 49.48 km auftreten, einerseits logarithmisiert, andererseits nicht logarithmisiert. Auch hier macht eine Logarithmisierung Sinn, da man tiefere y-Werte kaum mehr interpretieren könnte.

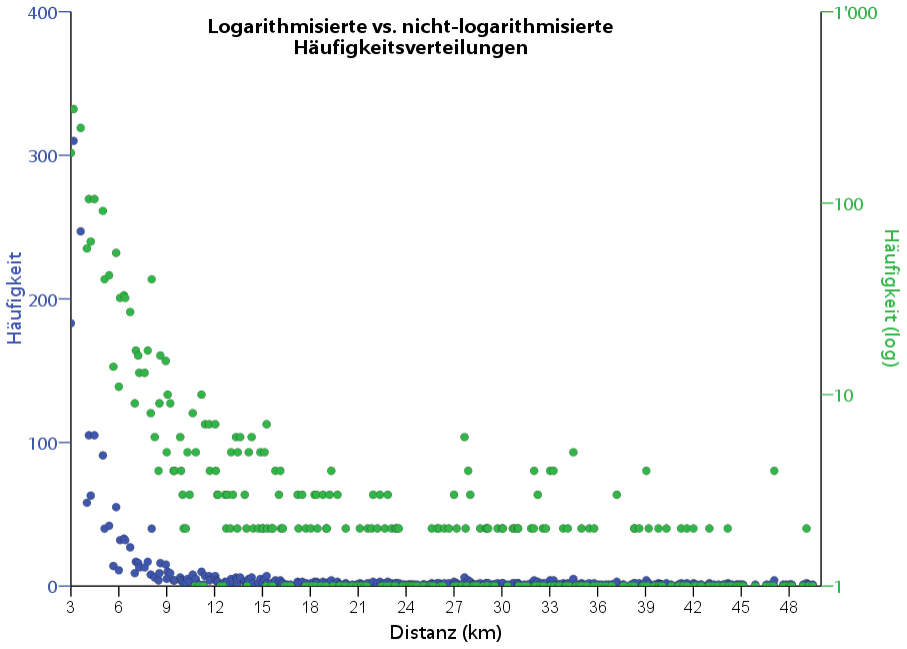


Abbildung 5.8: Häufigkeitsverteilung der Distanzen zwischen 3 und 49.48 km. Blau: nicht logarithmisierte Häufigkeit, grün: logarithmisierte Häufigkeit.

Bei der Häufigkeitsverteilung in Abbildung 5.9 wird die x-Achse in 5 km-Klassen eingeteilt. Man kann nun selbst ohne Logarithmisierung die y-Werte interpretieren, da in jede Klasse verhältnismässig etwa gleich viele Stichproben fallen. Für die Verteilung der Häufigkeiten der Distanzen zwischen 3 und 49.48 km ergibt eine explorative Datenanalyse weitere statistische Werte, welche in der Tabelle 5.6 einzusehen sind.

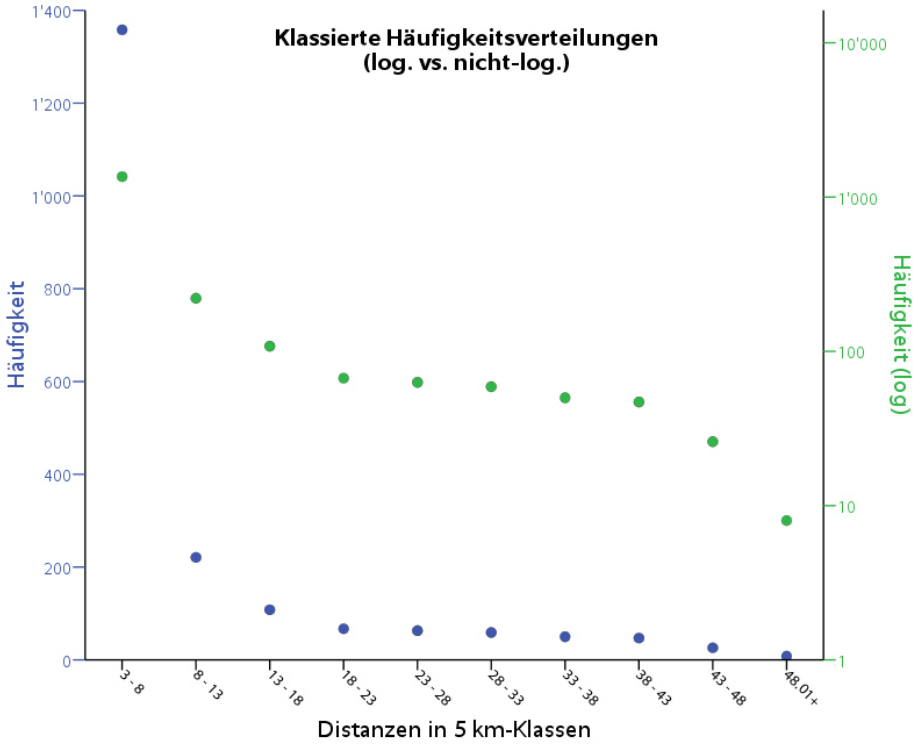


Abbildung 5.9: Klassierte Häufigkeitsverteilungen der Distanzen zwischen 3 und 49.48 km. Blau: nicht logarithmisierte Häufigkeit, grün: logarithmisierte Häufigkeit.

Tabelle 5.6: Statistische Kenngrössen der Distanzen von 3 bis 49.48 km.

Anzahl Proben	2'190
Mittelwert	9.32 km
Median	5 km
Standardabweichung	10.06 km

5.2.5 Interpretation

Die beiden Dichtefunktionen in der Abbildung 5.3 zeigen die Verteilungen der Stichproben. Die linke Karte zeigt alle Aufnahmeorte, welche von GEOGRAPH gesammelt wurden. Dagegen sieht man in der rechten Abbildung bloss die Verteilung jener Stichproben, welche die Beziehung A *near* B enthalten. Es treten zwar in ganz Grossbritannien Stichproben auf, diese sind jedoch sehr ungleichmässig verteilt. Es ist eine Verlagerung der Dichte in einigen Städten zu beobachten: Sobald nur noch „A *near* B"-Beziehungen vorherrschen, nimmt die Dichte der Stichproben

in Städten ab und gleichzeitig entstehen neue Dichtezentren ausserhalb von Städten.

Werden alle gültigen Distanzen in einer Häufigkeitsverteilung betrachtet, wird ersichtlich, dass die meisten Stichproben im Bereich gegen 0 km und nicht in der Nähe von 49.48 km zu liegen kommen. Alle Mediane, ausser jener von der Verteilung 3 bis 49.48 km, weisen eine Distanz von 1 km auf, was bedeutet, dass jeweils über 50% aller Distanzen 0 und 1 km darstellen. Der Mittelwert von der Verteilung 0 bis 2.83 km liegt bei 1.06 km, derjenige von der Verteilung 0 bis 49.48 km ist 2.28 km. Die unterschiedliche Grösse des Spektrums und die im Vergleich geringe Differenz zwischen den Mittelwerten untermauern ebenfalls den Fakt, dass sich die allermeisten Distanzen in der Anwendung mit *near* in einem sehr schmalen Bereich abspielen. Zwischen 3 und 49.48 km treten bloss noch 14.7% der Stichproben auf, der Median liegt hier bei 5 km, was wiederum eine sehr starke Abnahme der Häufigkeiten im Zusammenhang mit einer Zunahme der Distanz widerspiegelt.

Betrachtet man die Verteilung in 5.4, gelangt dieser eben beschriebene Modus der kleinen Distanzen auch visuell zum Ausdruck, dabei muss die Selektions-Inhomogenität berücksichtigt werden. Um dieser Inhomogenität Abhilfe zu leisten, werden in der Abbildung 5.5 die Distanzen klassiert. Somit sind nur noch 11 verschiedene gleich grosse Distanzklassen vorhanden, statt 666 verschiedene Distanzen. Obwohl viele Distanzen somit zusammengefasst werden, lässt sich ebenfalls derselbe Trend erkennen: Je höher die Distanzen sind, desto tiefer ist die Häufigkeit. Lediglich zwischen 20 und 30 km lässt sich ein leichter Anstieg der Häufigkeiten erkennen.

Eine logarithmisierte Darstellung aller Verteilungen sorgt dafür, dass auch höhere Stichproben voneinander unterschieden werden können. Lediglich bei der Verteilung der Distanzen von 0 bis 2.83 km ist keine Logarithmisierung nötig. Die Distanz von 1 km stellt mit 39.6% wiederum den absoluten Spitzenwert dar, bei 2.24 km ist gegenüber 2 km ein leichter Anstieg zu erkennen.

5.3 Regionale Unterschiede

Im vorliegenden Unterkapitel wird untersucht, ob *near* in verschiedenen Regionen in Grossbritannien unterschiedlich verwendet wird. Die Abbildung 5.10 gewährt eine Übersicht über den durchschnittlichen Gebrauch aller gültigen Distanzen vom Datenbestand der GEO-DB pro 100x100 km Kacheln in ganz Grossbritannien. Die Grundlage des 100x100 km Rasters bildet das bereits erläuterte Koordinatensystem von Ordnance Survey. Diese Übersicht erlaubt einen groben Eindruck über regionale Unterschiede in der Anwendung von *near*. In diesem Falle macht es Sinn, alle

Beziehungen bis 49.48 km miteinzubeziehen, da pro 100x100 km Kachel im Schnitt 270 Proben zu liegen kommen und somit grössere Distanzen nicht als Einzelfälle resp. Ausreisser gelten.

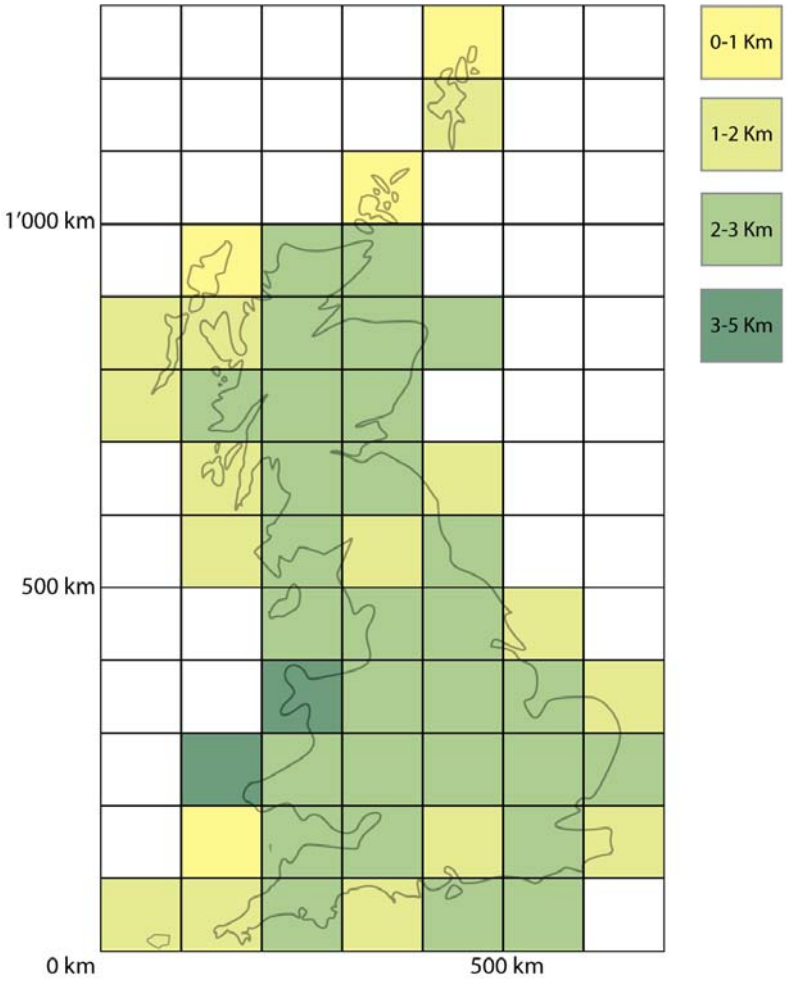


Abbildung 5.10: Durchschnittlich verwendete Distanzen bis 49.48 km in 100x100 km Kacheln.

5.3.1 Nord- und Süd-Grossbritannien

Um zu überprüfen, ob zwischen dem Norden und dem Süden von Grossbritannien ein Unterschied zwischen den verwendeten Distanzen besteht, werden im Norden wie auch im Süden je vier 100x100 km Kacheln näher untersucht. Es werden wiederum alle gültigen Distanzen von 0 bis 49.48 km betrachtet, da es sich doch um relativ grosse Stichproben handelt. Die Abbildung 5.11 zeigt die beiden Ausschnitte. Der nördliche Ausschnitt entspricht den Kacheln NS, NN, NH und NJ, sie enthalten insgesamt 1'368 Proben. Der südliche Ausschnitt wird von den Kacheln

ST, SU, SO und SP gestellt und umfasst 3'002 Beziehungen.

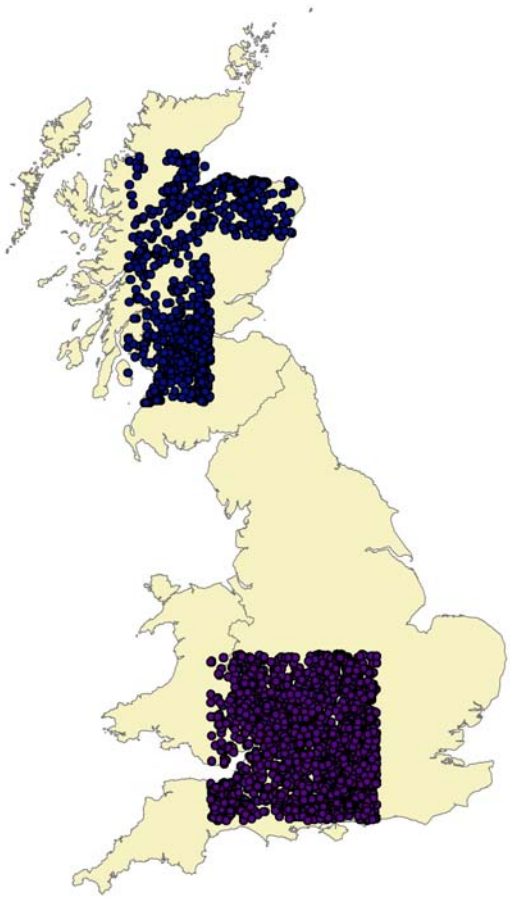


Abbildung 5.11: Für den Vergleich zwischen Nord- und Süd-GB werden je 4 100x100 km Kacheln ausgewählt. Die Punkte stellen die Aufnahmeorte der Fotos dar. $n_{norden} = 1'368, n_{sueden} = 3'002$.

Vergleicht man die Häufigkeitswerte für die Distanzen in diesen beiden Regionen, dann erhält man die in der Abbildung 5.12 dargestellten Diagramme. Die Häufigkeiten müssen einerseits logarithmisiert werden und andererseits müssen die Distanzen in 5 km Klassen eingeteilt werden, damit die Grafik interpretierbar wird. Die Klassierung ist nötig, da ab ca. 10 km in den meisten Fällen jeder Distanz nur noch eine Probe zugeordnet werden kann. Teilt man die x-Achse in 5 km Bereiche ein, so finden sich mehrere Proben, welche eine Häufigkeit von grösser als 1 haben. Die folgende Tabelle 5.7 zeigt die Unterschiede zwischen einigen statistischen Kenngrössen. Dabei kann festgestellt werden, dass im Norden im Zusammenhang mit *near* grössere Distanzen verwendet werden als im Süden.

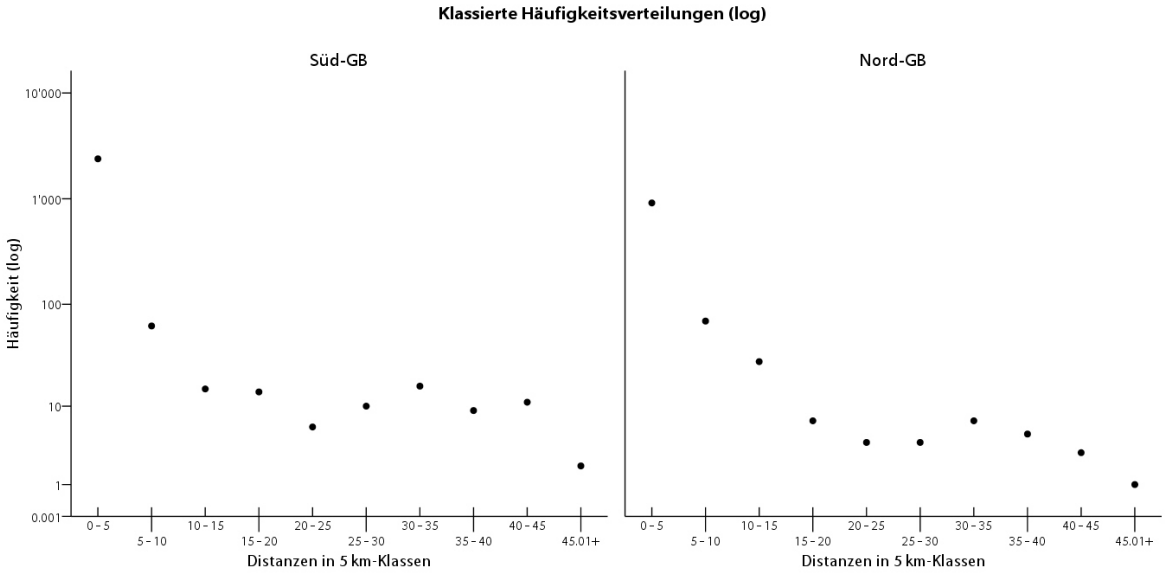


Abbildung 5.12: Vergleich der Häufigkeiten bezüglich der assoziierten Distanzen in je vier 100x100 km Kacheln im Süden und im Norden Grossbritanniens. Die Häufigkeitswerte sind logarithmisiert und die Distanzen in 5 km-Klassen unterteilt.

Tabelle 5.7: Statistische Kenngrössen für die Distanzen von 0 bis 49.48 km in den je 4 100x100 km Kacheln im Norden und im Süden Grossbritanniens.

	Nord-GB	Süd-GB
Anzahl Proben	1'368	3'002
Mittelwert	2.46 km	2.12 km
Median	1 km	1 km
Standardabweichung	4.86 km	4.78 km

Es wird zur Überprüfung, ob diese beiden Proben signifikant unterschiedlich sind, erneut ein Mann-Whitney-U-Test durchgeführt. Da $p > 0.05$ muss angenommen werden, dass sich die beiden Stichproben nicht signifikant voneinander unterscheiden.

5.3.2 Edinburgh und London

Um einen Vergleich anzustellen zwischen den Hauptstädten London und Edinburgh wird darauf geachtet, dass die Stichproben jeweils innerhalb der Stadtgrenze liegen. Daher werden für die beiden Städte zwei verschieden grosse Ausschnitte gewählt. Um die 260 km^2 grosse Fläche der 0.5 Millionen Einwohner Stadt Edinburgh abzudecken wird ein 20x20 km Raster aufgespannt, für die 1'500 km^2 grosse Fläche der 7.5 Millionen Einwohner Stadt London wird ein 40x40 km

Raster benutzt. Für London ergeben sich 79 Proben, für Edinburgh deren 11. Da es sich hierbei um relativ kleine Stichproben handelt, wurden lediglich Distanzen bis 11.31 km berücksichtigt. Die Tabelle 5.8 liefert eine Übersicht über die statistischen Werte.

Tabelle 5.8: Vergleich Edinburgh-London.

	Edinburgh	London
Anzahl Proben	11	79
Mittelwert	1.02 km	1.78 km
Median	1 km	1 km
Standardabweichung	0.87 km	1.7 km

Führt man mit diesen beiden Verteilungen einen Mann-Whitney-U-Test durch, erhält man einen p-Wert grösser als 0.05. Man kann also ebenfalls kein signifikanter Unterschied zwischen diesen beiden Verteilungen feststellen, obwohl zu erkennen ist, dass in London durchschnittlich grössere Distanzen verwendet werden. Berücksichtigt man nun die zuvor angestellten Aussagen über den Nord-Süd-Vergleich, drängt sich die Frage auf, weshalb in London höhere Werte vorhanden sind. Um die beiden Verteilungen graphisch miteinander vergleichen zu können werden zwei Wahrscheinlichkeitsverteilungen generiert, welche in der Abbildung 5.13 sichtbar sind.

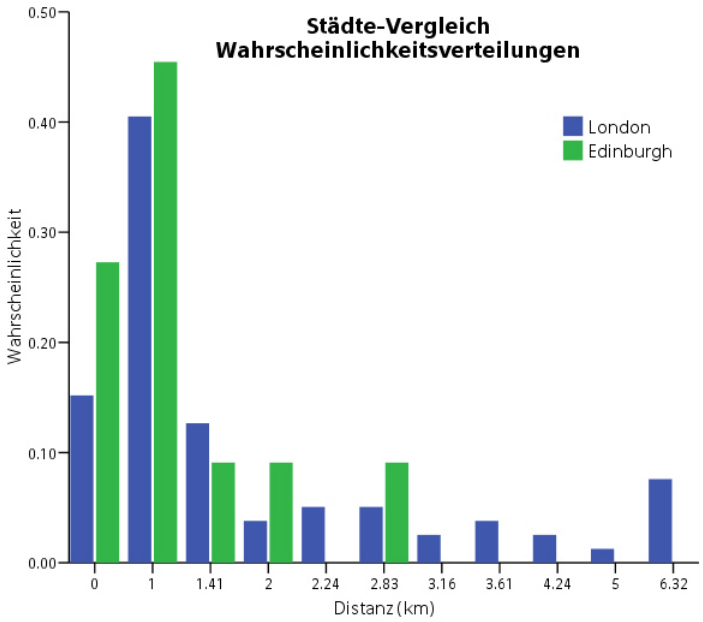


Abbildung 5.13: Wahrscheinlichkeitsverteilung von London und Edinburgh.

5.3.3 Stadt und Land

Wie unterscheiden sich die Distanzen in einer städtischen und in einer ländlichen Region? Für die städtische Region werden die bereits oben ermittelten Werte von London verwendet. Die ländliche Region wird repräsentiert durch einen 30x30 km grossen Ausschnitt mit einem Zentrum, welches rund 80 km westlich von London liegt. Dieses Zentrum wird durch ein kleines Dorf mit knapp 300 Einwohnern namens Peasemore repräsentiert, das weitgehend von landwirtschaftlich kultivierten Flächen umgeben ist. In der Geographie ist die Definition einer ländlichen Region bzw. einer Landschaft umstritten. Der gewählte Ausschnitt ist am ehesten einer Kulturlandschaft zuzuordnen, da sich das Landschaftsmerkmal „kultivierte Felder“ mehrmals wiederholt (Orgin, 1999; Hard und Gliedner, 1979). Das Satellitenbild in der Abbildung 5.14 von Google Earth zeigt den 900 km^2 grossen Ausschnitt um Peasemore.

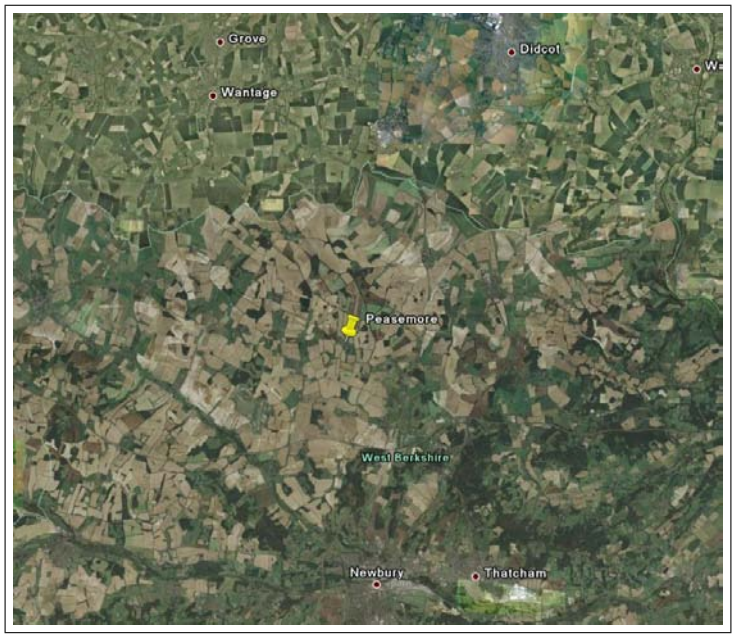


Abbildung 5.14: 30x30 km Ausschnitt, welcher eine ländliche Region repräsentiert. Quelle: Google Earth, 13.05.08

In der Tabelle 5.9 wird ein Vergleich zwischen den statistischen Werten aufgezeigt. Die Werte der ländlichen Region weisen einen tieferen Mittelwert als die Distanzen in der Stadt auf. Stellt man einen Mann-Whitney-U-Test mit den beiden Verteilungen an, um zu eruieren, ob die beiden Proben aus derselben Grundgesamtheit stammen oder nicht, führt dies dazu, dass man die Nullhypothese nicht verwerfen kann. Die Abbildung 5.15 liefert die Wahrscheinlichkeitsverteilungen für die verwendeten Distanzen von 0 bis 11.31 km in diesen beiden Regionen.

Tabelle 5.9: Vergleich Stadt – Land

	Stadt	Land
Anzahl Proben	79	240
Mittelwert	1.78 km	1.412 km
Median	1 km	1.414 km
Standardabweichung	1.7 km	0.92 km

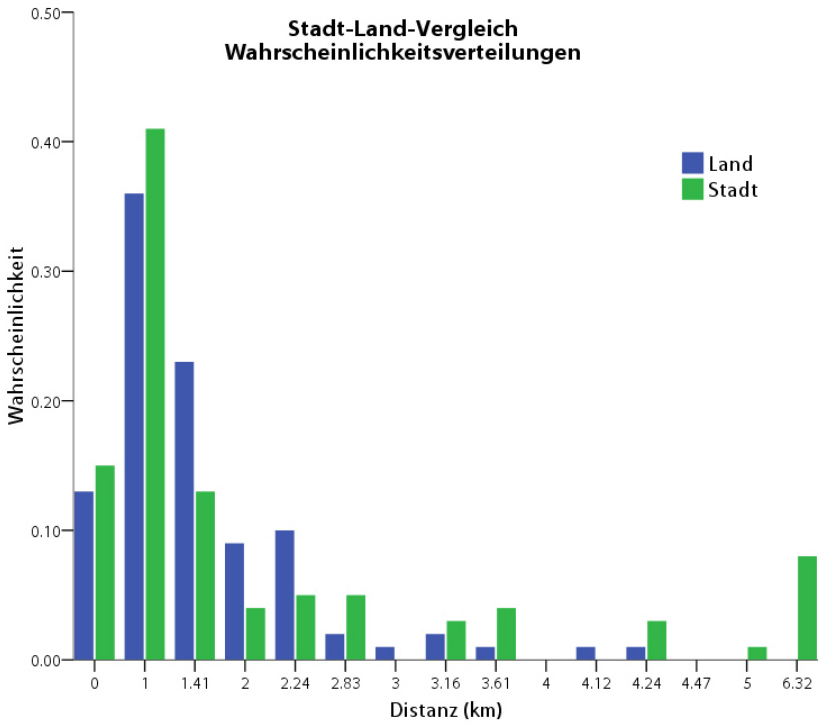


Abbildung 5.15: Vergleich der Distanzen zwischen Stadt und Land.

5.3.4 Schottisches Hochland und Englisches Flachland

Es wird untersucht, ob ein Unterschied zwischen den Highlands in Schottland und dem Flachland in Nordengland in Bezug auf die Anwendung von *near* feststellbar ist. Als Ausgangspunkt für das Schottische Hochland wird der höchste Berg Grossbritanniens ('Ben Nevis', 1'344 m.ü.M.) gewählt. Das Englische Flachland wird durch die Region um Leicester repräsentiert. Um diese beiden Zentren wird ein 40x40 km Ausschnitt aufgespannt. Alle darin enthaltenen Distanzen von 0 bis 11.31 km werden darauf ausgewertet. Die Tabelle 5.10 zeigt einen Vergleich zwischen den wichtigsten statistischen Kenngrössen. In der Abbildung 5.16 sind die unterschiedlichen Wahrscheinlichkeitsverteilungen ersichtlich. Ein Mann-Whitney-U-Test liefert die Erkenntnis, dass sich

die beiden Proben nur schwach signifikant unterscheiden. Der p-Wert liegt bei 0.018 und ist somit kleiner als das Signifikanzniveau von 0.05 ($p < 0.05$).

Tabelle 5.10: Vergleich Schottisches Hochland – Flachland

	Hochland	Flachland
Anzahl Proben	62	285
Mittelwert	1.88 km	2.27 km
Median	1 km	1.41 km
Standardabweichung	1.93 km	2.12 km

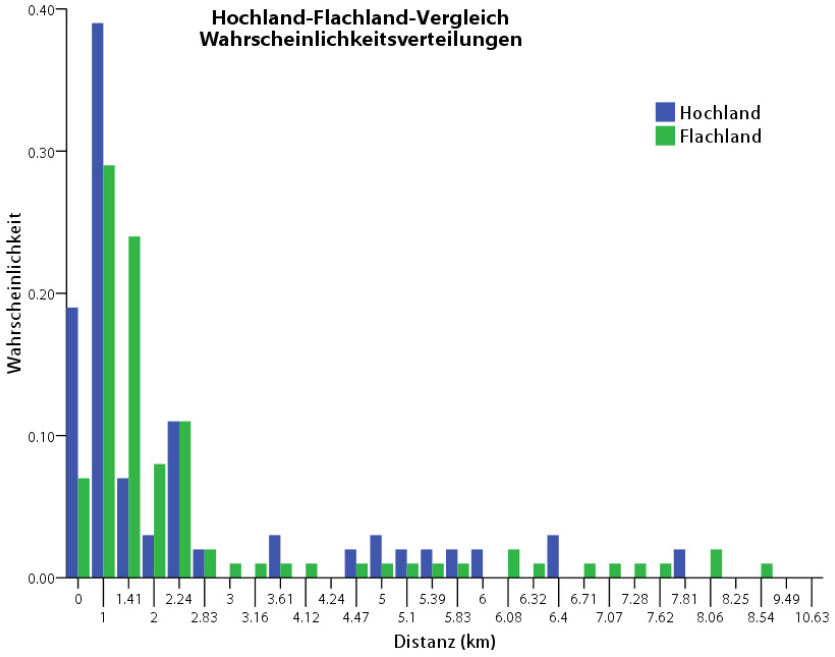


Abbildung 5.16: Das Schottische Hochland im Vergleich mit dem Englischen Flachland.

5.3.5 Interpretation

Die Abbildung 5.10 lässt erkennen, dass auf den Halbinseln und Inseln um Grossbritannien kleinere Distanzen verwendet werden im Zusammenhang mit *near*. Aufgrund dieses „Inseleffektes“ entstehen im Mittel Distanzen zwischen 0 und 2 km, wobei die durchschnittlich verwendeten Distanzen im Landesinneren zwischen 2 und 3 km liegen.

In Nord-Grossbritannien werden durchschnittlich 0.34 km grössere Distanzen verwendet als in Süd–Grossbritannien. Bei einem Signifikanzniveau von 0.05 stammen die beiden Stichproben

jedoch aus derselben Grundgesamtheit. Die Häufigkeitsverteilungen der beiden Regionen weisen daher eine sehr grossse Ähnlichkeit auf. Wiederum ist bei beiden Verteilungen ein Anstieg der Häufigkeiten zwischen 25 und 30 km zu erkennen. Im Städtevergleich verfügt die Stadt London mit 1.78 km über einen 0.76 km höheren Mittelwert als Edinburgh. In den Wahrscheinlichkeitsverteilungen ist auffällig, dass keine der 80 Proben über 6.32 km ist, obwohl alle Stichproben bis 11.31 km berücksichtigt werden. Bei 6.32 km ist der hohe Wert von London markant, welcher für eine relativ hohe Standardabweichung von 1.7 km sorgt. Bei beiden Städten ist der Median wie auch der Modus bei 1 km, die zweithäufigste Distanz ist jeweils 0 km.

Beim Stadt-Land-Vergleich kann festgestellt werden, dass der Mittelwert vom ländlichen Ausschnitt um 0.37 km leicht tiefer liegt als jener von der Stadt. Obwohl der ländliche Ausschnitt von 240 Stichproben repräsentiert wird, tritt keine Distanz über 5 km auf, auch wenn alle Distanzen bis 11.31 km berücksichtigt werden. Beim Vergleich zwischen dem Schottischen Hochland und dem Englischen Flachland tauchen dagegen Distanzen bis 10.63 km auf, weshalb die Stichproben stark gestreut sind. Dies drückt sich in einer deskriptiven Datenanalyse in der verhältnismässig hohen Standardabweichung aus.

5.4 Korrelationen mit der Distanz

5.4.1 Distanz – Einwohnerzahl

Distanzen von 0 bis 49.48 km

Es wird die Fragestellung behandelt, ob eine Korrelation zwischen der Distanz der Koordinate A und B und der Einwohnerzahl des Toponyms B erkennbar ist. Dafür werden nur noch diejenigen Toponyme benutzt, welche in der Einwohnertabelle von Grossbritannien eingetragen sind. Diese Toponyme stellen daher ausschliesslich Dörfer und Städte dar. Durch dieses Vorgehen reduzieren sich die insgesamt 14'856 Beziehungen auf 3'209. Die Tabelle 5.11 zeigt auf, wie viel mal welche Toponyme resp. Städte mindestens 10-mal verwendet werden. Die zweithäufigst verwendete Stadt „Keith“ wirft die Frage auf, ob es sich hiermit tatsächlich um die Nordschottische Stadt „Keith“ mit 4'520 Einwohnern handelt oder eher um einen Personennamen, was zu falschen Resultaten führen würde.

Tabelle 5.11: Städte und Dörfer, die mindestens 10-mal verwendet wurden.

Stadt	Häufigkeit	Stadt	Häufigkeit	Stadt	Häufigkeit
Leicester	44	Coldingham	14	Cirencester	11
Keith	35	Llangollen	14	Marlborough	11
Stafford	27	Louth	14	Milton	11
Spalding	20	Upton	14	Stranraer	11
Wakefield	20	Grantham	13	Carlton	10
Wymondham	19	Sheffield	13	Crook	10
Falkirk	16	Forfar	12	Dunbar	10
Garstang	16	Portpatrick	12	Tywyn	10
Melton Mowbray	16	Portsoy	12	Ware	10
Saltash	16	Sevenoaks	12	Watford	10
Barnard Castle	14	Chester	11		

Wenn man den Mittelwert der Distanzen von 0 bis 49.48 km betrachtet, lässt sich feststellen, dass sich dieser von 2.28 auf 3.51 km erhöht. Die Tabelle 5.12 liefert eine Zusammenstellung dieser unterschiedlichen Resultate.

Tabelle 5.12: Städte und Dörfer implizieren grössere Distanzen, als die restlichen Toponyme.

	Alle	Ohne Städte und Dörfer	Nur mit Städten und Dörfer
Mittelwert	2.28 km	1.94 km	3.51 km
Median	1 km	1 km	2.24 km
SD	4.89 km	4.85 km	4.87 km
Anzahl Proben	14'861	11'652	3'209

Trägt man alle 3'209 Distanzen von 0 bis 49.48 km gegenüber den Einwohnerzahlen ab, dann erhält man das Streudiagramm in Abbildung 5.17. Die Punktreihe, welche sich auffällig auf der Höhe der y-Achse von 330'574 gebildet hat, sind allesamt Fotografien mit Kommentaren, welche *near Leicester* enthalten. Die Punktreihe auf der Höhe von 439'866 Einwohnern widerspiegelt *near Sheffield*. Da die Stichproben von GEOGRAPH nicht normalverteilt sind, wird für die Berechnung des Korrelationskoeffizienten auf die nichtparametrische Spearman'sche Rangkorrelation zurückgegriffen. Ein zweiseitiger Test gibt einen Wert von $\rho = 0.396$ zurück. Dies bestätigt den visuellen Eindruck, dass die verwendeten Distanzen mit *near* keinen Zusammenhang haben mit der Einwohnerzahl einer Stadt.

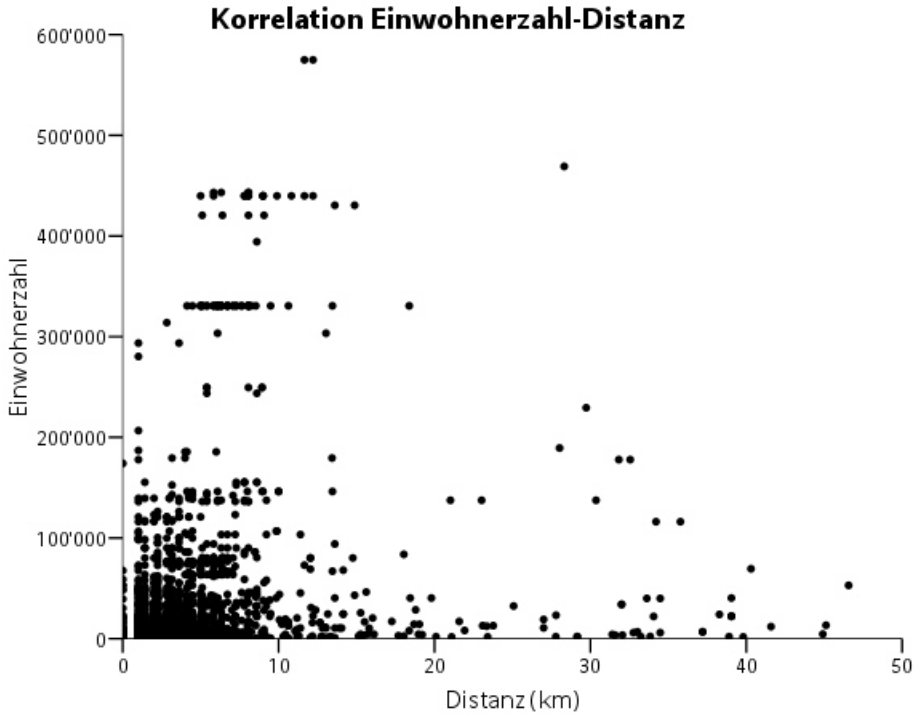


Abbildung 5.17: Einwohnerzahlen von Toponymen gegenüber den verwendeten Distanzen bis 49.48 km. $n = 3'209, \rho = 0.396$.

Distanzen von 0 bis 11.31 km

Von den insgesamt 14'683 Distanzen von 0 bis 11.31 km weisen 3'079 Beziehungen ein Toponym auf, welches in der Einwohnertabelle von Grossbritannien mit einer Einwohnerzahl attributiert ist. Ein Diagramm, welches diese Stichproben gegenüber den entsprechenden Einwohnerzahlen plottet, findet sich in der Abbildung 5.18.

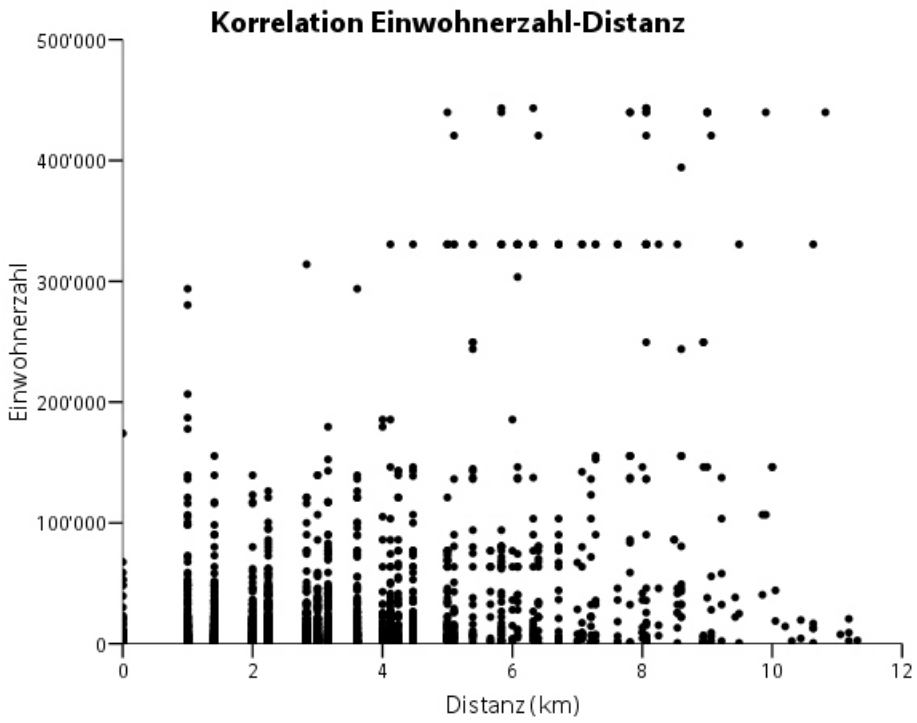


Abbildung 5.18: Einwohnerzahlen von 3'079 Toponymen gegenüber den verwendeten Distanzen bis 11.31 km. $n = 3'079, \rho = 0.395$.

Regionale Unterschiede

Nachdem ganz Grossbritannien auf einen Zusammenhang zwischen den Einwohnerzahlen und den verwendeten Distanzen überprüft wurde, wird nun untersucht, ob diesbezüglich ein Unterschied zwischen dem Norden und dem Süden Grossbritanniens festzustellen ist. Es werden wiederum jeweils dieselben vier 100x100 km Kacheln verwendet, wie im Kapitel 5.3.1. Berücksichtigt man die Distanzen von 0 bis 11.31 km, weisen im Norden von total 1'315 Proben 452 ein Toponym mit einer korrespondierenden Einwohnerzahl auf. Im Süden sind dies deren 573 von insgesamt 2'923 Beziehungen. In der Abbildung 5.19 werden die beiden Streudiagramme in einem Quervergleich abgebildet. In der Tabelle 5.13 werden die relevanten statistischen Grössen bezüglich der verwendeten Distanzen im Norden wie auch im Süden aufgezeigt.

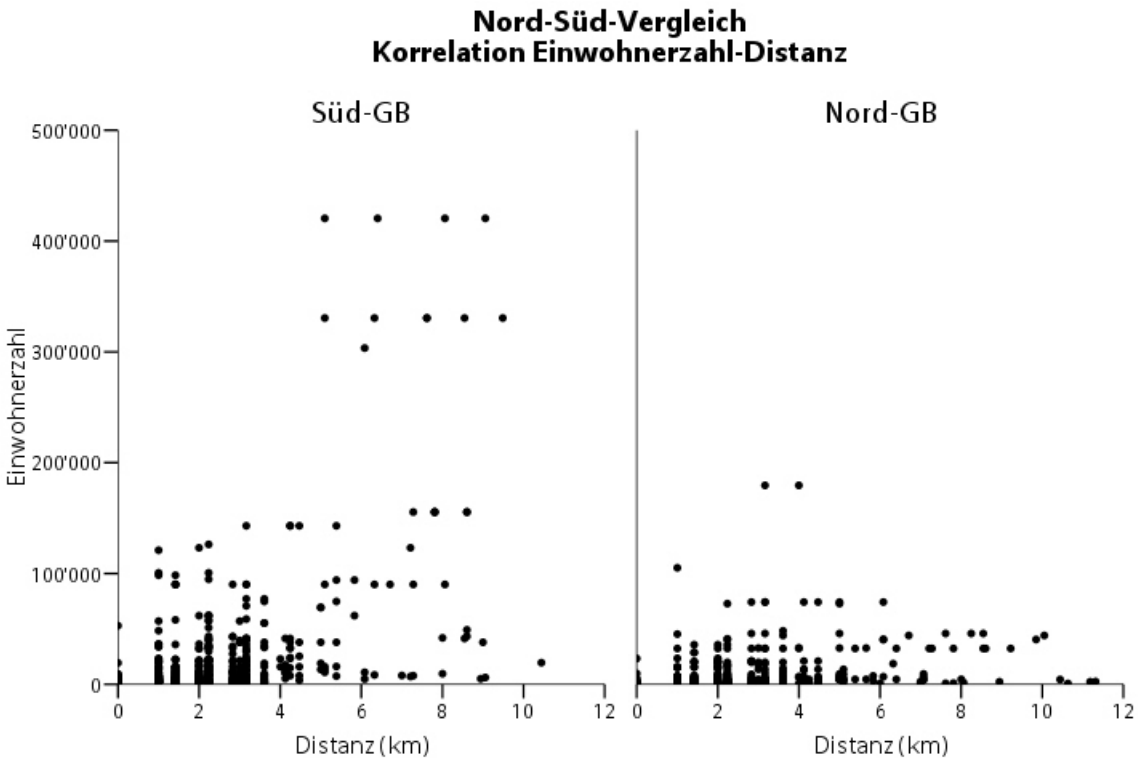


Abbildung 5.19: Einwohnerzahlen vom Toponym gegenüber der verwendeten Distanz bis 11.31 km in Nord- und Süd-GB, $n_{nord} = 452, \rho = 0.381. n_{sued} = 573, \rho = 0.477$.

Tabelle 5.13: Statistische Werte für N- und S-GB.

	Nord-GB	Süd-GB
Anzahl Proben	452	573
Mittelwert	2.70 km	2.42 km
Median	2.24 km	2.00 km
Standardabweichung	1.99 km	1.78 km

5.4.2 Distanz – Imageclass

Die Fotografen von GEOGRAPH mussten jeder Fotografie aus einer vorgegebenen Auswahl eine *Imageclass* zuordnen. Diejenigen Bildklassen, welche mindestens 100-mal verwendet wurden, finden sich in der Tabelle 5.14. Dabei kristallisieren sich 29 verschiedene Klassen heraus. Es wird untersucht, ob ein Zusammenhang zwischen den verwendeten Distanzen und den *Imageclasses* besteht.

Tabelle 5.14: Imageclasses, welche mindestens 100-mal verwendet wurden.

Objekt	Häufigkeit	Objekt	Häufigkeit	Objekt	Häufigkeit
Farmland	1572	Bridge	255	House	126
Farm	566	Woodland	248	Bridleway	123
River	417	Cottages	224	Road scene	120
Field	380	Road Junction	218	Valley	109
Road	375	Countryside	167	Rural View	107
Fields	338	Moorland	148	Railway bridge	103
Farm buildings	296	Farmhouse	145	Pasture	100
Track	261	Coastline/Beaches	141		
Country road	259	Railway	135		
Footpath	258	Canal	132		
Lane	257	Barn	129		

In der Abbildung 5.20 werden alle diese Klassen gegenüber der jeweils mittleren Distanz, welche bei den mindestens 100 Beziehungen auftreten, eingetragen. Im Zusammenhang mit der *Imageclass* „Barn“ wurden die geringsten Distanzen verwendet, mit der Bezeichnung „Moorland“ tauchen deutlich die grössten Distanzen auf.

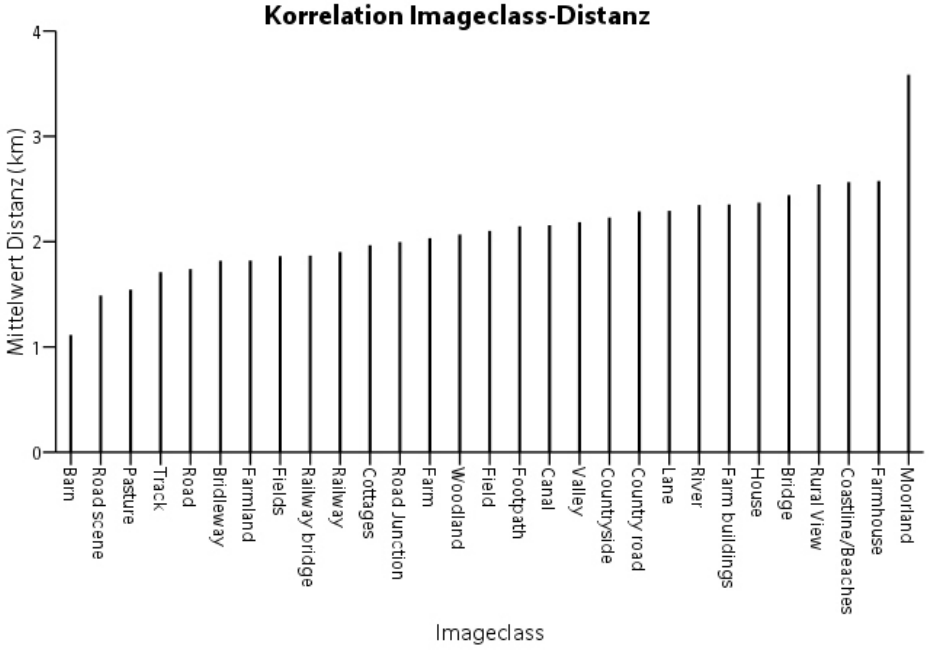


Abbildung 5.20: Imageclasses, die in GEOGRAPH mindestens 100-mal verwendet werden.

5.4.3 Interpretation

Beziehungen, die sich auf eine Stadt oder ein Dorf beziehen, werden mit einer Distanz von 3.51 km verwendet. Falls man genau diejenigen Beziehungen weglässt und sich auf die restlichen Toponyme konzentriert, dann sinkt der Mittelwert auf 1.94 km. Sobald also Städte bzw. Dörfer als Bezugsobjekt gewählt werden, kann man davon ausgehen, dass sich die Distanz erhöht, im Gegensatz dazu, wenn man sich auf sonstige Toponyme wie zum Beispiel Felder, Bäche, Plätze und dergleichen bezieht. Am meisten wurde die 330'000-Einwohnerstadt Leicester im Zusammenhang mit *near* verwendet, gefolgt vom 5'000-Einwohnerstädtchen Keith im Norden Schottlands. Erstaunlicherweise tauchen Grossstädte wie Edinburgh und London nicht unter den Städten auf, welche mindestens 10 Mal verwendet wurden: *near Edinburgh* tritt 3 Mal, *near London* 2 Mal in der Datenbank von GEOGRAPH auf.

In der Abbildung 5.17 kann festgestellt werden, dass keine klare Korrelation zwischen der Einwohnerzahl und der verwendeten Distanz zu erkennen ist. Wie zu erwarten ist, tauchen im Süden deutlich höhere Einwohnerzahlen auf, die nie mit einer Distanz kleiner als 5 km verwendet werden. Auch im Norden werden die einwohnerreichen Städte jeweils erst ab einer Distanz von ca. 3 km verwendet. Dies bedeutet jedoch nicht, dass ab dieser Distanz keine kleineren Städte mehr verwendet werden. Stellt man einen Nord-Süd-Vergleich an (vgl. Tab. 5.13, S. 57), kann festgestellt werden, dass wiederum im Norden grössere Distanzen verwendet werden. Es spielt also keine Rolle, ob man alle Toponyme oder nur jene, welche eine Stadt oder ein Dorf sind, auswertet - die grösseren Distanzen treten im Norden auf. Eine Korrelation ist jedoch ebenfalls nicht erkennbar. Bei der Verwendung der Imageclass und den entsprechenden Distanzen kann kein Muster erkannt werden. Es werden nicht unbedingt jene Imageclasses mit den grössten Distanzen verwendet, welche gefühlsmässig die grösste Ausdehnung haben (vgl. „Field“ vs. „Fields“).

5.5 Verteilung der Toponyme im Gazetteer

Verhaltet sich die Verteilung der Toponyme der Daten von GEOGRAPH ähnlich wie jene der Toponyme vom Gazetteer? Wäre dies der Fall, müsste man für weitere Untersuchungen berücksichtigen, dass ein Zusammenhang mit dem Gazetteer besteht und dass nicht unbedingt ausschliesslich unterschiedliche menschliche Wahrnehmung die Distanzen definieren.

Um diesen Sachverhalt zu analysieren wird in jedem Datensatz von GEOGRAPH der Koordinate A ein zufällig gewähltes Toponym aus dem Gazetteer zugeordnet und danach die Distanz ausgerechnet. Dabei werden nur jene Einträge im Gazetteer berücksichtigt, welche sich auf denjenigen Raum beziehen, in welchem auch die Stichproben der GEO-DB auftreten. Dabei muss jedoch beachtet werden, dass um den Raum, den man für die Koordinaten des Gazetteers ein-

räumt, eine Pufferzone addiert werden muss. Die Pufferzone entspricht der maximalen Distanz, welche bei der Auswertung der Daten von GEOGRAPH festgelegt wird. Dies muss so sein, da ein Toponym der Daten von GEOGRAPH theoretisch ausserhalb des zu untersuchenden Raumes liegen kann, falls sich die Koordinate A in der Peripherie des Untersuchungsgebietes befindet. Dieser Versuch wird an drei Beispiel-Paaren durchgeführt. Es wird Nord- mit Süd-Grossbritannien, städtisches- mit ländlichem Gebiet und das Schottische Hochland mit dem Englische Flachland verglichen. Es ist zu erwarten, dass sich die zufällig generierten Beziehungen von der Stichprobe der GEO-DB signifikant voneinander unterscheiden. Ansonsten wären die von den Fotografen gewählten Toponyme zufällig gewählt worden.

5.5.1 Nord- und Süd-Grossbritannien

Im Kapitel 5.3.1 wurde der Unterschied zwischen Nord- und Süd-Grossbritannien im Zusammenhang mit der Verwendung von *near* aufgezeigt. Es konnte festgestellt werden, dass im Norden im Durchschnitt 0.34 km grössere Distanzen auftreten, berücksichtigt man alle Distanzen bis 49.48 km. Innerhalb desselben Nord-Ausschnittes, inklusive des oben erwähnten Puffers von 49.48 km (das entspricht einer zusätzlichen Fläche von rund 600 km^2), befinden sich 75'003 Toponyme, welche im Gazetteer registriert sind. Im südlichen Ausschnitt tauchen 81'453 von den insgesamt 258'978 Einträgen auf. Im Datenbestand von GEOGRAPH befinden sich im Norden 1'368 und im Süden 3'002 Aufnahmeorte. Mittels eines Zufallsgenerators werden den entsprechenden A-Koordinaten zufällig Toponyme aus dem Gazetteer mit deren B-Koordinaten zugeordnet. Die Tabelle 5.15 zeigt einen Vergleich zwischen den statistischen Kenngrössen der Distanzen vom Datensatz von GEOGRAPH und den zufällig entstandenen Distanzen im Norden und im Süden Grossbritanniens. In der Tabelle 5.16 werden die prozentualen Differenzen zwischen den beiden Mittelwerten der beiden Verteilungen aufgezeigt. Diese Darstellung vereinfacht es, eine allfällige Gemeinsamkeit bzw. eine Tendenz zwischen den beiden Datensätzen in den verschiedenen Regionen zu erkennen.

5.5.2 Stadt und Land

Im Kapitel 5.3.3 wurde ein Vergleich angestellt zwischen einer städtischen und einer ländlichen Region. Man konnte erkennen, dass in der Stadt im Mittel weitere Distanzen verwendet werden als auf dem Land. Nun wird wiederum untersucht, wie sich die zufällig generierten Beziehungen in den entsprechenden Ausschnitten verhalten. In der Tabelle 5.17 sind die Resultate aufgezeichnet. Die Tabelle 5.18 zeigt die prozentuale Differenz an, welche zwischen dem Datenbestand von GEOGRAPH und dem zufällig generierten Datensatz entsteht.

Tabelle 5.15: Vergleich der zufällig generierten Distanzen mit denen von GEOGRAPH (km).

		Anzahl	Mittelwert	Median	SD
Nord GB	Zufällig	1'368	548.40	548.79	103.21
	GEOGRAPH	1'368	2.46	1.00	4.86
Süd GB	Zufällig	3'002	451.52	450.82	85.46
	GEOGRAPH	3'002	2.12	1.00	4.78

Tabelle 5.16: Prozentualer Unterschied des Mittelwertes von Nord- zu Süd-Grossbritannien.

	Differenz
Zufällig	-17.67%
GEOGRAPH	-13.82%

5.5.3 Schottisches Hochland und Englisches Flachland

Im Kapitel 5.3.4 wurden die Beziehungen im Schottischen Hochland und im Englischen Flachland auf Flächen von je 40x40 km untersucht und es konnte festgestellt werden, dass im Flachland durchschnittlich grössere Distanzen verwendet werden bezüglich der Beziehung A *near* B. Nun werden ebenfalls zufällige Distanzen im entsprechenden Ausschnitt generiert und mit den Resultaten vom Datensatz von GEOGRAPH miteinander verglichen. Der Vergleich kann in den Tabellen 5.19 und 5.20 eingesehen werden.

Tabelle 5.19: Vergleich zwischen den Distanzen einer zufälligen Wahl von Toponymen aus dem Gazetteer und den Distanzen von GEOGRAPH in Bezug auf das Schottische Hochland und das Englische Flachland (km).

		Anzahl	Mittelwert	Median	SD
Hochland	Zufällig	62	25.2	24.13	11.13
	GEOGRAPH	62	1.88	1.00	1.93
Flachland	Zufällig	285	26	25.5	11.93
	GEOGRAPH	285	2.27	1.41	2.12

Tabelle 5.20: Prozentualer Unterschied von den Highlands zu den Lowlands.

	Differenz
Zufällig	+3.17%
GEOGRAPH	+20.74%

5.5.4 Interpretation

In der Tabelle 5.15, bei der der Norden mit dem Süden verglichen wird, ist zu erkennen, dass zwischen den zufällig generierten Distanzen und den Distanzen von GEOGRAPH erwartungsgemäss

Tabelle 5.17: Vergleich zwischen den Distanzen einer zufälligen Wahl von Toponymen aus dem Gazetteer und den Distanzen von GEOGRAPH in Bezug auf eine städtische Region und eine ländliche Region bei einer max. Distanz von 11.31 km (km).

		Anzahl	Mittelwert	Median	SD
Stadt	Zufällig	79	29.55	29.07	13.39
	GEOGRAPH	79	1.78	1.00	1.7
Land	Zufällig	240	21.29	20.85	10.11
	GEOGRAPH	240	1.412	1.414	0.92

Tabelle 5.18: Prozentualer Unterschied vom städtischem zu ländlichem Gebiet.

	Differenz
Zufällig	-27.95%
GEOGRAPH	-20.67%

ein grosser Unterschied vorliegt. Darüber hinaus kann man jedoch ein interessantes Verhalten beobachten: Beide Datenbanken weisen im Norden grössere Distanzen als im Süden auf. Prozentual verhält sich dieser Unterschied in etwa gleich. Ebenfalls beim Stadt-Land-Vergleich kann man feststellen, dass die Tendenz zwischen den beiden Ausschnitten ähnlich ist. Einzig beim Vergleich zwischen dem Schottischen Hochland und dem Englischen Flachland ist die Gemeinsamkeit nicht ganz so klar wie bei den vorherigen Vergleichen. Es ist jedoch ebenfalls bei der zufällig generierten Datenbank und der GEO-DB die gleiche Tendenz zu erkennen. Zusammenfassend kann gesagt werden, dass sich die durchschnittlichen Distanzen zweier Regionen innerhalb des Gazetteers relativ gleich verhalten wie die Distanzen in der GEO-DB.

Kapitel 6

Diskussion

Die folgende Diskussion beruht durchgehend auf der Annahme, dass sich die Ergebnisse immer auf Bildbeschriftungen stützen. Andersartige Verwendungen von *near* könnten durchaus zu anderen Resultaten führen. Vergleicht man die 1.1, kann festgestellt werden, dass Distanzen in dieser Grössenordnung im Zusammenhang mit Bildbeschreibungen praktisch auszuschliessen sind.

6.1 Bezug zur Literatur

Die dieser Diplomarbeit zugrunde liegende Datenbank basiert auf dem kognitiven Basiswissen, über welches die Teilnehmer von der sie umgebenden Umwelt verfügen. Egenhofer und Mark (1995) beschreiben dieses Wissen als „Naive Geography“, wobei sie aus dem Verständnis darüber einen intuitiveren Zugang zu GIS erhoffen. Einige Resultate in dieser Arbeit werden dieses Verständnis ebenfalls bereichern, wie zum Beispiel die später diskutierte Erkenntnis, dass kein signifikanter Unterschied darin besteht, ob man *near* im Zusammenhang mit einer einwohnerreichen oder -armen Stadt verwendet. Trifft man die Annahme, dass Städte mit vielen Einwohnern über eine grössere Fläche verfügen als Städte mit weniger Einwohnern, kann man daher den Vorschlag von Stevens und Coupe (1978), dass die Menschen sich räumliche Information hierarchisch merken, in Frage stellen. Würden Menschen räumliche Information hierarchisch strukturieren, dann hätten einwohnerreiche und somit flächenmässig grössere Städte logischerweise einen höheren Stellenwert. Dies würde dazu führen, dass grosse Städte mit kleineren und kleine Städte mit grösseren Distanzen verwendet werden müssten, da man Objekte mit einem höheren Stellenwert als weniger weit entfernt empfindet (Worboys, 2001).

Stevens und Coupe wurden unter anderen von Friedmann und Brown (2000) kritisiert. Sie behaupten, dass die räumliche Vorstellung vor allem durch nichträumliche Informationen gebildet

wird. Dazu gehören Faktoren wie Reisezeit, Reisekosten, Attraktivität des Zielortes, Aktivität am Zielort, usw., welche bereits in diversen Arbeiten untersucht wurden (Gahegan, 1995; Hernández et al., 1995; Guesgen, 1999; Yao und Thill, 2005). In dieser Arbeit werden jedoch bloss räumliche Faktoren behandelt, welche die Wahrnehmung der Distanz resp. die Wahrnehmung von *Nähe* beeinflusst. Zum Beispiel wird der Faktor „Einwohnerzahl" untersucht, der unter der oben erwähnten Annahme ebenfalls einen räumlichen Charakter impliziert. Ein weiterer räumlicher Faktor wird durch die Verschiedenheit der untersuchten Regionen gegeben. Über diesen Faktor wurde bislang keine Studie bezüglich der Wahrnehmung von *Nähe* publiziert, die sich auf ein länderübergreifendes Untersuchungsgebiet ausdehnte. Diese Arbeit hingegen untersucht Stichproben in ganz Grossbritannien. Dagegen hat Worboys (2001) auf einem Universitätscampus in England ein Experiment über *nearness* durchgeführt (vgl. Kapitel 2.2.1, S. 11). Dieses Experiment lässt jedoch bloss Schlüsse zu, wie die Wahrnehmung von *Nähe* auf lokaler Ebene verwendet wird. Im Vergleich zu dieser Arbeit muss erwähnt werden, dass die Bilddatenbank von GEOGRAPH mit 5'345 Teilnehmern um ein Tausendfaches mächtiger ist und daher mit Grossbritannien auch ein entsprechend grosses Untersuchungsgebiet abdecken kann.

Nach Guesgen (1999) wandeln Menschen quantitative Information oft in qualitative Werte um, um den Sinn einer Information besser zu verstehen. In dieser Arbeit ist es genau umgekehrt: Der Teilnehmer macht eine qualitative Einschätzung einer Distanz (z.B. „This is a farm *near* Barmouth.") und diese wird anhand der korrespondierenden Koordinaten in eine quantitative Information (Distanz in Kilometer) gewandelt. Diese quantitative Information wird ausgewertet und in Diagrammen graphisch dargestellt. Die Diskussion dieser Diagramme wird teilweise wiederum qualitativ formuliert, wie zum Beispiel die Erkenntnis, dass im Zusammenhang mit grösseren Städten nicht unbedingt grössere Distanzen verwendet werden müssen.

6.1.1 Einteilung des Raumes

Es existieren verschiedenste Klassierungen des Raumes (vgl. Kapitel 2, S. 6). Um den Untersuchungsraum in dieser Arbeit zu beschreiben, eignet sich die Klassifikation nach Montello (1993). Er teilt den Raum in die vier auf der Seite 7 aufgelisteten Klassen ein. Es existiert nicht bloss eine einzige Klasse, der man den gesamten Inhalt dieser Arbeit zuteilen könnte. Vielmehr sind Parallelen zwischen den einzelnen Klassen und der Einteilung des Raumes in dieser Arbeit erkennbar. Die Tabelle 6.1 zeigt einen Vorschlag, wie man den Inhalt dieser Arbeit und die Klassen von Montello verbinden könnte. Für die Klasse „Figural" findet sich jedoch kein passender Vergleich. Da von GEOGRAPH vorgegeben ist, dass Fotos die charakteristische Geographie repräsentieren müssen, ist es unmöglich, einen Bildausschnitt zu wählen, der kleiner als ein Mensch ist.

Tabelle 6.1: Zuordnung der Raumklassen von Montello mit den verschiedenen Ebenen.

Montellos Klassen	Zuordnung
Figural	
kleiner als ein Mensch	→ kein Vergleich
Vista	
von einem Standort aus sichtbar	→ Bildinhalt der Fotografien
Environmental	
Fortbewegung benötigt, um alles zu sehen	→ Bewegungsraum einzelner Teilnehmer
Geographical Space	
Karte nötig	→ ganzes Untersuchungsgebiet

6.1.2 Referenzrahmen

Die im vorangehenden Unterkapitel erwähnten räumlichen Einteilungen implizieren Referenzrahmen (spatial frame of reference) mit unterschiedlichen Skalen. Die Skala bzw. der Kontext ist entscheidend, wenn man Untersuchungen über räumliche Beziehungen anstellt. Im Experiment von Worboys an der Keele University wird die Entfernung von der „Library“ zur „Barnes Hall“ von allen Teilnehmern als eine grosse Distanz eingestuft. Dieselbe euklidische Distanz im Untersuchungsgebiet dieser Arbeit würde jedoch in die kleinste Distanzklasse fallen. Diese Studie stützt sich auf zwei Referenz- bzw. Bezugsrahmen: Der erste entspricht dem gesamten potentiellen Untersuchungsgebiet selbst (Grossbritannien), der zweite (ein Subreferenzrahmen) definiert sich durch den Bewegungsraum und die Kenntnisse über den Raum der einzelnen Teilnehmer. Dieser entspricht dem individuellen Kontext jedes Teilnehmers. Es kann angenommen werden, dass die Teilnehmer in ihrem Bewegungsraum oft ortskundig sind, daher kann man davon ausgehen, dass der individuelle Referenzrahmen bei allen Teilnehmern in etwa gleich gross ist. Wären die meisten Bilder zum Beispiel in den Ferien (in unbekannteren Regionen) entstanden, dann würden die Teilnehmer möglicherweise „mental maps“ bilden, welche einen Referenzrahmen vom Ferien- bis zum Wohnort umfassen und somit die ganze Wahrnehmung der Distanzen verzerren könnte. Dies bedeutet, dass die Grössenordnung einer „mental map“, welche die Ausdehnung des Referenzrahmens definiert, einerseits vom Standort der Teilnehmer und andererseits vom kognitiven Basiswissen über die Umgebung abhängig ist. Um die Wichtigkeit dieses Referenzrahmens bzw. des Kontextes zu veranschaulichen, soll auf das Beispiel in Abbildung 1.1, Seite 2 hingewiesen werden.

Es könnte durchaus sein, dass sich die Menschen in Grossbritannien nicht überall in einem gleich grossen Gebiet gut auskennen und somit auch ortsabhängig verschieden grosse individuelle Referenzrahmen existieren. Daher werden Regionen miteinander verglichen, um herauszufinden,

ob der individuelle Referenzrahmen regional bzw. landschaftsbedingt ist und somit grössere oder kleinere Distanzen hervorbringen könnte. Denn ist der individuelle Referenzrahmen tendenziell gross, wird erwartet, dass auch grössere Distanzen im Zusammenhang mit *near* verwendet werden. In dieser Arbeit kann der individuelle Referenzrahmen jedoch nicht direkt ermittelt werden, daher besteht bloss die Möglichkeit anhand der ausgewerteten Distanzen rückwirkend auf den zugrunde liegenden Referenzrahmen zu schliessen. In Kapitel 6.3.1 wird über den Nord-Süd-Unterschied diskutiert. Da im Norden im Mittel grössere Distanzen verwendet werden, könnte angenommen werden, dass die Einwohner von Nord-Grossbritannien über einen grösseren individuellen Referenzrahmen verfügen als jene vom Süden Englands.

6.1.3 Asymmetrie

Sainsbury (1995) hat sich mit der Frage beschäftigt, ab wann viele Sandkörner als ein Sandhaufen bezeichnet werden können (vgl. Kapitel 2.2, Seite 8). Sinngemäss dasselbe Problem stellt sich bei der Frage, ab welcher Distanz man sich *nahe* einer Stadt befindet. Diese Grenze kann bei der Analyse der GEO-DB gut ermittelt werden. Dabei muss das Mittel aller Distanzen genommen werden, die im Zusammenhang mit einer Stadt verwendet werden. Dagegen ist es nicht möglich, die folgende Frage zu beantworten: Entstehen unterschiedliche Distanzen, je nach dem, ob man sich von einer Stadt wegbewegt oder ob man sich auf die Stadt zubewegt? Dasselbe Prinzip findet sich bei der Sandhaufen-Frage: Entsteht dieselbe Grenze, ob man den Sandhaufen aufbaut oder ob man ihn abbaut? Falls dabei nicht dasselbe Resultat entsteht (und somit richtungsabhängig ist) dann spricht man von Asymmetrie.

Um die Asymmetrie mathematisch zu erfassen, entwickelte Worboys (1996) den Begriff der „relativen Distanzen“. Diese relativen Distanzen legen nebst dem Referenzrahmen bzw. dem Kontext verschiedener Städte auch die asymmetrischen Entfernungen der Städte zueinander fest (siehe Kapitel 2.2, Seite 11). Da bei Worboys zentral gelegene Städte eine geringere relative Distanz aufweisen, müsste man erwarten, dass die Städte in Südengland im Zusammenhang mit *near* ebenfalls mit geringeren Distanzen verwendet werden. Dem ist tatsächlich auch so (vgl. Kapitel 5.3.1, S.46). Weshalb dies so sein könnte, wird später in diesem Kapitel in der Diskussion der Resultate ergründet.

Da die Koordinate von A und B jeweils eine Kachel von 1 km^2 repräsentiert und somit meistens mehrere Toponyme pro Koordinate vorhanden sind, ist eine eindeutige Zuordnung des korrekten Toponyms A nicht möglich. Würde man neben der Beziehung A *near* B auch die Beziehung B *near* A vorfinden, könnte man wenigsten aussagen, dass es sich um eine symmetrische Beziehung handelt. Dies schliesst jedoch nicht aus, dass die Beziehung zusätzlich einen asymme-

trischen Charakter besitzen kann: Es könnte sein, dass *near* in einem Bereich angewendet wird und dieser muss nicht gleich gross sein für beide Richtungen. Wo genau die Grenzen liegen, kann man daher nicht herausfinden. Eine Untersuchung einer allfälligen Asymmetrie zwischen zwei Punkten in den Daten der GEO-DB ist aus diesen Gründen leider nicht möglich.

6.2 Häufigkeitsverteilungen der Distanzen

6.2.1 Test auf zufällige Verteilung

Vergleicht man in einem 5x5 km Raster eine zufällige Verteilung mit derjenigen von der GEO-DB (vgl. Abb. 5.2, S. 34), ist es offensichtlich, dass diese beiden Stichproben unabhängig voneinander sind. Der p-Wert ist grösser als 0.05 und bestätigt somit diesen Eindruck. Die Teilnehmer von GEOGRAPH wählen im Zusammenhang mit *near* in dem Fall nicht einfach ein willkürliches Toponym B, sondern eines, das tatsächlich relativ nahe beim Aufnahmeort liegt. Daher entspricht dieses Resultat den Erwartungen. Würden die beiden Verteilungen ähnlich sein (was bedeuten würde, dass die Stichproben der GEO-DB zufällig zu Stande gekommen wären), würde es keinen Sinn machen, diese Stichproben auszuwerten.

Bei der zufälligen Verteilung wie auch bei derjenigen der GEO-DB ist bei der Distanz von 2.24 km ein Anstieg der Häufigkeiten ersichtlich (siehe Abbildung 5.2). Dies bedeutet, dass die Eintrittswahrscheinlichkeit nicht für alle Distanzen gleich hoch ist. Der Grund für diesen starken Anstieg bei 2.24 km liegt nicht bei der häufigeren Verwendung dieser Distanz, sondern muss auf die ungleichen Flächenanteile pro Distanz in einem Raster zurückgeführt werden (vgl. Tabelle 5.1). Auf Grund dieser Begebenheit stellt sich die Frage, ob eine Normalisierung der Daten sinvoller wäre, dies könnte den unterschiedlichen Flächenanteil quasi neutralisieren. Dabei müssten die Häufigkeiten der Distanzen so manipuliert werden, als würde jede Distanz in einem Raster gleich viel Flächenanteil besitzen. Eine normalisierte zufällige Verteilung würde demnach alle Distanzen über dieselbe Häufigkeit verfügen. Da eine Normalisierung jedoch die originalen Werte der Stichproben stark verändert, ist es sinnvoll, dies zu unterlassen. Der Vorteil dieser Entscheidung liegt darin, dass die Resultate die unverfälschten Stichproben repräsentieren, wie diese in die Datenbank importiert wurden und somit keine allfällig unerwünschten und unkontrollierten Umrechnungseffekte auftreten können. Der Nachteil besteht allerdings darin, dass einem der Fakt der unterschiedlichen Flächenaufteilung bei der Auswertung der Resultate stets bewusst sein muss.

6.2.2 Geographische Verteilung der Stichproben

In Abbildung 5.3 sind zwei Dichtefunktionen zu sehen, wobei die linke Funktion alle 362'308 Aufnahmeorte des ganzen Datensatzes von GEOGRAPH repräsentiert. Die rechte Dichtefunktion zeigt eine Teilmenge aller Aufnahmeorte, nur jene 14'861 Stichproben, in deren Datensätzen die Beziehung A *near* B auftritt. Die beiden Dichtefunktionen unterscheiden sich insbesondere darin, dass sich oft die Dichte der Häufigkeiten der Stichproben von den Stadtzentren weg auf ländlichere Gebiete verlagert, wie dies zum Beispiel bei London gut zu beobachten ist. Bildkommentare, die A *near* B enthalten, treten offensichtlich weniger oft in der Stadt auf als alle übrigen Bildkommentare. Den gleichen Vorgang kann man in Edinburgh verfolgen: Auf die höchste Dichte in der linken Abbildung folgt eine äusserst geringe Dichte der Stichproben.

Diese Verlagerung kann möglicherweise darauf zurückgeführt werden, dass man sich in einer Stadt praktisch immer in unmittelbarer (Sicht-)Nähe eines Toponyms befindet, wie zum Beispiel *auf* einem Platz oder *in* einem Park oder *in* einer Stasse. In einer Stadt werden daher möglicherweise häufiger Ausdrücke wie *at, in, on, beside, by, to the left/right* verwendet, welche oft eine Distanz von praktisch Null implizieren. Sobald die Dichte der Toponyme abnimmt und man sich öfters auf ein Toponym referenzieren muss, das sich nicht just am Aufnahmeort befindet, dann werden für eine Standortbeschreibung des öftern räumliche Relationsbegriffe wie *close to, not far, nearby, thereabout, around* und natürlich *near* verwendet, welche mit einer gewissen Distanz verwendet werden können.

6.2.3 Distanzen von 0 bis 49.48 km

Die Abbildung 5.4 zeigt die Verteilung der Häufigkeiten aller gültigen Stichproben, welche im Zusammenhang mit *near* verwendet wurden. 97% der Beziehungen weisen eine kleinere Distanz als 10 km auf. Die relativ kleine Standardabweichung deutet in Anbetracht des Mittelwertes ebenfalls auf eine Konzentration der Stichproben um 2 bis 3 km hin. Da es sich bei diesen Stichproben ausschliesslich um Beziehungen handelt, welche die Wendung A *near* B enthalten, wäre ein komplett anderes Bild auch nicht erklärbar. Mit dieser Verteilung kann nun die Vermutung bzw. die intuitive Annahme bestätigt werden, dass man im Zusammenhang mit *near* durchschnittlich relativ kleine Distanzen verwendet.

Bei einer Distanz von etwa 30 km erfährt die Häufigkeit einen leichten Anstieg. Dies ist darauf zurückzuführen, dass ein und derselbe Teilnehmer am selben Ort mehrere Fotos getätigt hat und dabei in den Bildkommentaren immer dieselbe *near*-Toponym-Kombination angewendet hat.

Obwohl dies nur vier mal der Fall ist, fällt dies auf, da Stichproben in diesen Distanzbereichen nur noch vereinzelt auftreten. Auch eine Klassierung ändert das Bild nicht besonders stark, der Anstieg bei 30 km ist immer noch erkennbar. Weshalb dies so ist, kann mit den vorhandenen Daten nicht erklärt werden. Statistisch betrachtet unterscheiden sich diese „Ausreisser“ ebenfalls nicht signifikant von den anderen Stichproben.

6.2.4 Distanzen von 0 bis 11.31 km

Im Bereich von 0 bis 11.31 km kann erneut die starke Abnahme der Stichproben mit wachsender Distanz beobachtet werden. In diesem Ausschnitt finden sich 96% der Stichproben bereits in den ersten 5.39 km. Daraus kann abgeleitet werden, dass allgemein sehr kleine Distanzen im Zusammenhang mit *near* verwendet werden, ob nun der beobachtete Ausschnitt, wie vorher besprochen, 49.48 km oder bloss 11.31 km gross ist. Dies erkennt man auch am Median, der bei beiden Verteilungen mit 1 km gleich ist, was bedeutet, dass mindestens 50% der Stichproben den Wert 0 oder 1 km hat. Dagegen schlagen sich die vereinzelt grossen Werte in der Verteilung bis 49.48 km im Mittelwert nieder: Die Verteilung von Distanzen bis 11.31 km hat mit 1.51 km einen 0.77 km kleineren Durchschnitt als die Verteilung von Distanzen bis 49.48 km. Beachtet man die sehr grosse Differenz zwischen den beiden untersuchten Distanzspektren, ist der Unterschied der beiden Mittel marginal, was wiederum auf die hohe Konzentration an Distanzen im Bereich zwischen 1 und 3 km zurückzuführen ist.

6.2.5 Distanzen von 0 bis 2.83 km

Wie oben erwähnt spielt sich die Gesamtheit der Stichproben im Mittel um 2.28 km ab. Daher macht es Sinn, das Distanzspektrum zwischen 0 und 2.83 km zu fokussieren, um das Verhalten der „grossen Masse“ zu untersuchen. Interessanterweise beziehen sich die Fotographen in ihren Bildkommentaren am meisten auf ein Toponym, das in einer der vier direkt angrenzenden Kacheln liegt (39.6%) und somit eine Distanz von 1 km im Zusammenhang mit *near* verwenden. Erst am zweithäufigsten werden Toponyme verwendet, die in derselben Kachel liegen (23.4%) und damit die Distanz 0 km ausdrücken. Im Zusammenhang mit *near* wird demnach nicht die kürzeste Distanz von 0 km am häufigsten verwendet, sondern eine Distanz von 1 km. Die dritthäufigste Distanz entsteht, wenn sich das Toponym B in einer der vier diagonal angrenzenden Kacheln befindet (19.1%). Dies entspricht einer Distanz von 1.41 km. Daraus kann man ableiten, dass sich diese Entfernung bereits nicht mehr so stark anbietet, um *near* zu verwenden. Eine klare Grenze in der Anwendung von *near* liegt jedoch erst ab einer Distanz von 3 km vor: Die zwei Distanzen 2.24 und 2.83 km werden zusammen nur noch in 11.1% der Fälle in diesem Spektrum verwendet. Die Erkenntnis, dass sich die meisten Distanzen innerhalb von 0 bis 2 km abspielen,

lässt schliessen, dass die Auflösung der Koordinaten in der Datenbank, welche bei 1 km^2 liegt, zu tief ist, um eine differenziertere Analyse durchzuführen. Würde man *near* hauptsächlich mit Distanzen von 30 km verwenden, dann könnte diese Auflösung reichen, um bereits schwache Veränderungen festzustellen.

Es könnte sein, dass in einer anderen Region auf der Erde im Zusammenhang mit *near* wesentlich grössere Distanzen verwendet werden. Zum Beispiel ist Russland mit einer Fläche von 17'000'000 km^2 rund 70 mal grösser als Grossbritanninen (230'000 km^2); Kanada, USA oder China verfügen ebenfalls über eine ca. 30-fach grössere Ausdehnung. Diese enorme Ausdehnung impliziert ebenfalls einen grösseren Referenzrahmen, der möglicherweise dazuführen könnte, dass Personen in diesen Staaten *near* im Zusammenhang mit grösseren Distanzen verwenden, als im Untersuchungsgebiet dieser Arbeit. Zudem ist in diesen weitläufigen Regionen die Toponym-Dichte oft sehr gering. Daher könnte es durchaus sein, dass eine Auflösung von 1 km^2 für eine differenziertere Analyse reichen könnte. Frank (1996); Hernández et al. (1995) erwähnen, dass auch verschiedene Kultur- und Sprachregionen unterschiedliche Wahrnehmungen von der Distanz bewirken könne.

Bei der Verteilung der Distanzen von 3 bis 49.48 km, in deren Spektrum bloss 14.7% aller gültigen Proben auftreten, ist der hohe Mittelwert (9.32 km) und der hohe Median (5 km) auffällig. Obwohl das Spektrum praktisch gleich gross ist, wie jenes, das alle Datensätze berücksichtigt, bestehen in diesen statistischen Kennwerten enorme Unterschiede. Dies ist ein weiterer Hinweis dafür, dass in Grossbritannien im Zusammenhang mit *near* sehr oft Distanzen bis 2.83 km verwendet werden.

6.3 Regionale Unterschiede

Die Übersichtskarte 5.10, die jeweils das Mittel der Distanzen einer 100x100 km Kachel repräsentiert, widerspiegelt unverkennbar den Mittelwert von 2.28 km: Die Mehrheit der Felder entspricht der Klasse 2 bis 3 km. In keiner Kachel werden im Schnitt grössere Werte ermittelt als 3 bis 5 km, obwohl bei dieser Darstellung alle gültigen Distanzen bis 49.48 km berücksichtigt wurden. Auffallend ist, dass ein „Inseleffekt“ zu erkennen ist: Praktisch auf allen Inseln und Halbinseln sind geringere Distanzen zu verzeichnen als in Kacheln, die im Landesinneren liegen. Im Mittel werden in diesen peripheren Gebieten Distanzen von 1 bis 2 km verwendet, einzelne Regionen fallen sogar in die Klasse 0 bis 1 km. Weshalb entsteht dieser Inseleffekt? Die Inseln sind im Verhältnis zum Festland sehr klein und die Toponyme liegen relativ eng beieinander. Auf einer

Insel wird es offensichtlich bevorzugt, sich auf Toponyme zu beziehen, die sich auch auf der Insel selber befinden und nicht auf dem Festland liegen. Für Fotografen auf einer Insel scheint demnach der Referenzrahmen nicht ganz Grossbritannien zu umfassen, sondern er beschränkt sich auf den Umfang der Insel. Dementsprechend kleiner ist daher auch der individuelle Referenzrahmen, der für diese kürzeren Distanzen verantwortlich ist. Es kann daher festgehalten werden, dass die physikalischen Begebenheiten einen starken Einfluss auf die Wahrnehmung der Distanzen haben.

Augenfällig sind auch die grossen Distanzen, die in Nord-Wales zum Ausdruck kommen. Die Verteilung in dieser Kachel verfügt mit 7.1% über verhältnismässig viele Werte mit Distanzen von 2.24 km. Ebenfalls taucht die Distanz von 3.16 km mit 1.9% relativ oft auf. Betrachtet man die einzelnen Bildkommentare genauer, dann stellt man fest, dass von den insgesamt 38 Beziehungen, welche eine Distanz von 2.24 km darstellen, 15 von ein und derselben Person zu verzeichnen sind. Es sind jedoch alles verschiedene Bildkommentare, die aus verschiedenen Beziehungen bestehen. Lediglich zwei Toponyme kommen doppelt vor, diese haben jedoch unterschiedliche Standorte sowie Kommentare. Es besteht daher kein Grund, diese Stichproben als ungültig zu erklären, obwohl es sich hierbei im Prinzip um einen Bias (systematischer Fehler) handelt. Der erhöhte Wert bei 3.16 km wurde vorwiegend von verschiedenen Teilnehmern generiert und weist mehrheitlich unterschiedliche Standorte und Toponyme B auf. Ein Grund, weshalb in Nord-Wales erhöhte Distanzen bevorzugt werden, ist, dass in Wales eine geringere Dichte an Toponymen als in England auftritt, was tendenziell weitere Distanzen zur Folge hat. Dies kann jedoch nicht der einzige Grund sein, da zum Beispiel in Schottland eine noch geringere Dichte vorhanden ist als in Wales. In Schottland werden jedoch (ausgenommen auf den Inselgruppen) keine grösseren Distanzen verwendet als in Wales. Daher müsste dieses Verhalten mittels in dieser Arbeit nicht untersuchten Faktoren, welche die Wahrnehmung der Distanzen verzerren können, vollständig erklärt werden (Gahegan, 1995; Hernández et al., 1995; Guesgen, 1999; Yao und Thill, 2005).

6.3.1 Vergleich verschiedener Regionen

Beim Vergleich zwischen dem Norden und Süden Grossbritanniens kann festgestellt werden, dass im Norden durchschnittlich grössere Distanzen verwendet werden als im Süden. Nun stellt sich die Frage, weshalb dies so sein könnte. Der Norden verfügt über mehr weite und unbewohnte Landschaften als der Süden, was einer geringeren Dichte der Toponyme entspricht. In dem 40'000 km^2 grossen Ausschnitt befinden sich im Norden 3.3 Toponyme pro km^2, in Süden sind es deren 7.3 pro km^2. Dies bedeutet, dass die Einwohner von Schottland eine andere Distanzwahrnehmung aufweisen als die Engländer. Schotten empfinden eine Distanz von beispielsweise 3 km weniger weit als Engländer: In England hat es innerhalb von 3 km mehr potentielle Ziele bzw. Toponyme als in Schottland. Daher ist in England die Wahrscheinlichkeit höher, dass man

sich auf ein näher gelegenes Toponym bezieht. Da man *near* in Relation zur Umwelt und zum Referenzrahmen verwendet (Gahegan, 1995), führt dies dazu, dass in Schottland *near* für weitere Strecken angewendet wird als in England.

Beim Vergleich zwischen Edinburgh und London stellt sich heraus, dass in London um 0.76 km weitere Distanzen verwendet werden. Obwohl dies den klarsten Unterschied aller Vergleiche bedeutet, können die beiden Proben nicht als signifikant unterschiedlich erklärt werden ($p > 0.05$). Beide Hauptstädte weisen über eine etwa gleich grosse Toponym-Dichte auf. Dass man in London *near* mit grösseren Distanzen verwendet, ist zurückzuführen auf die unterschiedliche Ausdehnung der Städte: London mit 1'500 km^2 verfügt im Gegensatz zu Edinburgh mit 260 km^2 über eine viel grössere Fläche. Ein Teilnehmer in London hat einen grösseren individuellen Referenzrahmen als ein Teilnehmer von Edinburgh und verwendet daher durchschnittlich grössere Distanzen im Zusammenhang mit *near*.

Auffällig ist jedoch beim Balkendiagramm in Abbildung 5.13, dass in London bei 6.32 km ein Anstieg der Häufigkeit vorliegt. Bei genauerem Betrachten der entsprechenden Datensätze erkennt man, dass es sich hierbei um denselben Teilnehmer handelt, der in seinen Kommentaren jeweils dieselbe A *near* B Beziehung verwendete. Da diese Datensätze das in der Methodik beschriebene Selektionsverfahren überstanden haben, besteht kein Grund, diese Datensätze zu ignorieren. Es stellt sich die Frage, ob eine Beschränkung, dass bloss zwei Bilder von derselben Person pro Quadratkilometer akzeptiert würden, sinvoller gewesen wäre. Lässt man im Nachhinein von den 6 deren 5 Werte weg, so verringert sich wohl der Mittelwert von 1.78 km auf 1.48 km, trotzdem ist der Unterschied zwischen Edinburgh und London immer noch präsent. Dieses Kriterium würde daher keine bemerkenswerten Auswirkungen auf das Resultat haben.

Analog zum Vergleich zwischen Nord- und Süd-Grossbritannien würde man auf Grund der unterschiedlichen Dichte der Toponyme erwarten, dass man in einer ländlichen Region weitere Distanzen verwendet als in der Stadt. Dies ist jedoch im Vergleich zwischen der ländlichen Region um Peasemore und London nicht der Fall (vgl. Kapitel 5.3.3, S. 50): In Städten herrscht eine hohe Toponym-Dichte, daher sind 1:50'000-Karten stark generalisiert. Viele kleine Toponyme werden weggelassen, stattdessen werden wenige, dafür grössere bzw. wichtigere Toponyme auf der Karte platziert (Baumgartner, 1990). Da grössere Toponyme weiter auseinander liegen als kleinere Toponyme, entstehen in der Stadt grössere Distanzen, obwohl die Dichte höher ist als auf dem Land. Sobald also ein Grad einer Toponym-Dichte erreicht ist, entsteht ein „Umkehr-Effekt“, so dass nicht mehr nur in Regionen mit einer tiefen Toponym-Dichte grosse Distanzen verwendet werden, sondern ebenfalls in Gebieten, in denen eine sehr hohe Toponym-Dichte herrscht (wie

z.B. in einer Stadt). Dieser Umkehr-Effekt tritt ein, sobald die reale Dichte der Toponyme höher ist als die Dichte der Toponyme, die auf der 1:50'000 Karte abgebildet ist. Sobald sich also der Fokus auf ein sehr dicht besiedeltes Untersuchungsgebiet richtet, entscheiden kartographische Regeln über die Distanz in Bezug auf *near* im Zusammenhang mit den Ergebnissen in dieser Arbeit. Das 1:50'000-Gazetteer hat keinen lückenlosen „Zugang" zu dicht besiedeltem Gebiet.

Im Englischen Flachland verwendet man grössere Distanzen im Zusammenhang mit *near* als im Schottischen Hochland. Dies kann darauf zurückzuführen sein, dass in hügeligen resp. gebirgigen Regionen oft auf ein Toponym referenziert wird, das nicht unbedingt eine Stadt ist, da in diesen Regionen die Städtedichte geringer ist als im Flachland. Im Flachland hat es jedoch eine höhere Städtedichte und somit auch mehr A *near* B Beziehungen, bei denen B eine Stadt ist. Wie wir in der Tabelle 5.12 auf Seite 54 erkennen können, werden im Zusammenhang mit Städten grössere Distanzen verwendet als mit den restlichen Toponymen. Weshalb dies so ist, wird im Kapitel 6.4.1 diskutiert.

Bei allen Vergleichen wurde das Distanzspektrum bis 11.31 km beobachtet (ausser beim Nord-Süd-Vergleich). Es ist daher auffällig, dass von den insgesamt 409 untersuchten Distanzen keine über 6.32 km liegt. Weshalb dies so sein könnte, ist mit den vorliegenden Informationen nicht möglich zu beantworten. Auf jeden Fall kann festgehalten werden, dass die konzeptionelle Distanz (im Gegensatz zur euklidischen) regional schwanken kann. Verantwortlich dafür sind die Toponym-Dichte, der Insel-Effekt (Referenzrahmen) und/oder der Umkehr-Effekt.

6.4 Korrelationen mit der Distanz

6.4.1 Korrelation Distanz-Einwohnerzahl

Wie oben erwähnt, werden grössere Distanzen im Zusammenhang mit *near* verwendet, wenn das Toponym eine Stadt oder ein Dorf ist, als wenn es sich um ein anderes Toponym handelt. Dies ist darauf zurückzuführen, dass das Toponym „Stadt" meistens über eine grössere Fläche als ein Platz, ein Park, eine Wiese, eine Strasse usw. verfügt. Daher beeinflusst die räumliche Ausdehnung der Stadt die Wahrnehmung der Distanzen zu einer Stadt: Man empfindet zum Beispiel eine Entfernung von 2 km zu einer Stadt als weniger weit, als dieselbe Entfernung zu einem kleinen Weiler, da der Mensch einer Stadt automatisch einen grösseren Referenzrahmen zuordnet. Diese Erkenntnis unterstützt die Hypothese, dass mit grösseren Städten weitere Distanzen verwendet werden.

Im Kapitel 2.4 wird die Fragestellung vorgestellt, ob ein Zusammenhang zwischen der Distanz und der Einwohnerzahl zu erkennen ist. Im Diagramm in Abbildung 5.17 wurden alle 3'209 Di-

stanzen eingetragen, die in einer Stichprobe im Zusammenhang mit einer Stadt oder einem Dorf verwendet wurden. Es ist ganz klar keine Korrelation zu erkennen. Die Einwohnerzahl von einer Stadt hat demnach keinen Einfluss auf die implizierte Distanz im Zusammenhang mit der Verwendung von *near*. Einwohnerreiche Städte müssten auf Grund ihrer grösseren Fläche mit weiteren Distanzen verwendet werden als weniger einwohnerreiche Städte (zu Stevens und Coupe (1978) bezüglich hierarchischem Merken von räumlichen Informationen, Kapitel 6.1, S. 63). Es finden sich aber trotzdem bei jeder Distanz Stichproben, die im Zusammenhang mit kleinen Dörfern entstanden sind. Grosse Städte wie Leicester und Sheffield werden mit einer beschränkten Distanz erwähnt. Die grösste Distanz im Zusammenhang mit Sheffield liegt bei 18 km. Die Stichproben ab 30 km werden nur noch mit Einwohnerzahlen verwendet, die unter 200'000 liegen. Eine Erklärung dafür, dass auch kleine Dörfer mit grossen Distanzen verwendet werden, könnte sein, dass einzelne kleine Dörfer einen hohen Bekanntheitsgrad haben und damit eine Wichtigkeit erlangen können, die höher ist als gewisse einwohnerreiche Städte (Yao und Thill, 2005). Dies erklärt jedoch noch nicht, weshalb ab 30 km keine Grossstädte mehr erwähnt werden.

Fokussiert man beim Diagramm in Abbildung 5.17 auf die ersten 11.31 km, kann wiederum festgehalten werden, dass keine Korrelation zwischen der Distanz und der Einwohnerzahl vorzufinden ist (siehe Abb. 5.18). Gut erkennbar ist, dass die allermeisten Distanzen unter 10 km liegen, danach treten nur noch spärlich Stichproben auf. Eine weitere Erkenntnis ist, dass die beiden Grossstädte Leicester und Glasgow erst ab einer Distanz von 4 km erwähnt wurden. Der Grund ist folgender: Jedem Städtenamen im Gazetteer wird diejenige Koordinate zugeordnet, welche das Stadtzentrum abdeckt. Bei grossen Städten heisst das, dass mit diesen Koordinaten bloss der Stadtkern mit einer Fläche von 1 km^2 abgedeckt wird. Liegt der Standort A nun zum Beispiel 4 km vom Zentrum entfernt, befindet man sich immer noch in der Stadt. Es ist daher naheliegend, dass man dann nicht sagt, dass man *in der Nähe* der Stadt ist, wenn man sich *in* der Stadt selber befindet. Ist man nun 1 km ausserhalb der Stadt von der Stadtgrenze entfernt, dann ist es durchwegs möglich, dass ein Teilnehmer seinen Standort als *nahe* zur Stadt definiert. Auch wenn er nun 1 km von der Stadt entfernt ist und daher den Standort A als nahe zur Stadt beschreibt, wird die Distanz zur Koordinate im Stadtzentrum berechnet. Dies würde in diesem Beispiel eine Distanz von 1 + x km bedeuten, wobei x der Distanz vom Stadtzentrum bis zum Stadtrand entspricht. Distanzen, die auf diese Art und Weise entstehen, können dem sogenannten „Zentrums-Effekt" zugeschrieben werden. Obwohl dieser Sachverhalt im Zusammenhang mit grossen Städten zu Fehlinterpretationen führen kann, muss gesagt werden, dass dieses Verhalten für sämtliche Beziehungen mit Grossstädten zu beobachten ist. Dies untermauert daher die Herleitung dieses Verhaltens und spricht für die Konsistenz der Daten.

Im Vergleich zwischen Nord- und Süd-Grossbritannien und den verwendeten Distanzen im Zusammenhang mit der Einwohnerzahl kommt zum Ausdruck, dass im Süden klar die grösseren Städte vorzufinden sind als im Norden. Wiederum stellt man fest, dass die grossen Städte nie mit einer Distanz von kleiner 3 bzw. 5 km verwendet wird. Der Grund dafür kann dem oben beschriebenen Effekt zugeordnet werden. Eine Korrelation zwischen der Einwohnerzahl und der Distanz ist nicht erkennbar. Es besteht also kein Zusammenhang zwischen der Einwohnerzahl von B und der Distanz von A zu B.

6.4.2 Ambiguität von Personennamen

Die Tabelle 5.11 listet alle Städte und Dörfer auf, die mindestens 10 Mal im Zusammenhang mit *near* verwendet wurden. Auffällig dabei ist, dass Keith (ein 4'520 Einwohner-Städtchen im Nordosten Schottlands), das mit 35 Erwähnungen am zweitmeisten vorkommt, vorwiegend als Männername bekannt ist. Daher stellt sich die Frage, ob zumindest teilweise A *near* Keith verwendet wurde, im dem Sinne, dass das Sujet auf dem Bild in der Nähe einer Person namens Keith liegt. Nach Überprüfung der einzelnen Datensätze kann anhand des Kontextes der Bildkommentare festgestellt werden, dass immer die Stadt Keith gemeint wurde und nicht eine Person. Unter anderem aus folgendem Grund: In der noch nicht selektionierten Datenbank konnte keine A *near* Keith Beziehung gefunden werden, bei der A nicht in unmittelbarer Nähe zur Stadt Keith liegt. Es kann durchaus sein, dass Bildkommentare verfasst wurden, bei der ein Personenname als Bezugsobjekt gewählt wurde. Da jedoch die Datenbank von GEOGRAPH moderiert wird, wurden diese Bilder nicht akzeptiert, da ein Kriterium darin besteht, dass sich weder der Bildinhalt noch der Bildkommentar auf Personen beziehen darf. Der Befund der Überprüfung dieser kritischen Beziehungen kann als Indiz dafür betrachtet werden, dass die Moderation von GEOGRAPH sehr gut funktioniert. Zur Veranschaulichung seien fünf solche kritische Bildkommentare aufgeführt (Koordinate Keith NJ4250):

- This is a registered knackery *near* Keith. (Koordinate A: NJ4149)
- This is an isolated derelict steading at Backpark *near* Keith. (Koordinate A: NJ3852)
- Berrylees Cottages and Farm *near* Keith. (Koordinate A: NJ4652)
- Cattle grazing on a cold Monday morning at bonnetill Farm *near* Keith. (Koordinate A: NJ4553)
- View towards the North from the bridge over Burn of Drum *near* Keith. (Koordinate A: NJ4551)

Weshalb also wurde trotzdem so oft die Stadt Keith verwendet? Bis auf zwei von insgesamt 35 Beziehungen wurden alle durch ein und dieselbe Person generiert. Sämtliche Kommentare sind jedoch unterschiedlich. Der Teilnehmer hat jedesmal ein anderes Sujet auf dem Bild, womit jede Beschreibung intuitiv zu Stande gekommen sein muss. Weiter wurde kontrolliert, ob dieser Teilnehmer ausschliesslich *near* in seinen Kommentaren verwendet hat, was die Validität dieser Datensätze in Frage stellen könnte: Insgesamt hat er 439 Bilder auf GEOGRAPH geladen (und gehört somit zu den fleissigsten Teilnehmern). Die Person hat in 33 Beiträgen *near* verwendet, was ein moderates Verhältnis ist (vgl. Kapitel 3.1, S. 17). Auf Grund der Moderation von GEOGRAPH, unter den gegebenen Kriterien und der eben beschriebenen Analyse dieser Datensätze besteht kein Grund, diese Beziehungen zu ignorieren.

Ein weiteres Beispiel einer Zweideutigkeit stellt der Name „Crook“ dar. Einerseits handelt es sich um einen Personennamen, andererseits um vier verschiedene Dörfer in Grossbritannien. Von den insgesamt zehn Erwähnungen beziehen sich neun auf die Stadt in der Provinz „Durham“ im Nordosten Englands. Wiederum werden die Datensätze kontrolliert und abgewogen, ob es sich um eine Person oder um die Ortschaft Crook handelt. Erneut kann festgestellt werden, dass sich sämtliche Kommentare auf die Ortschaft beziehen und nicht auf eine Person, die sich zufälligerweise in der Nähe des Aufnahmeortes befindet.

6.4.3 Korrelation Distanz-Imageclass

Bei der Auswertung der am 100 meisten verwendeten Imageclasses fällt der hohe Wert von 1’572 Zählern bei „Farmland“ auf. Die zweithäufigste Imageclass „Farm“ wird 566 mal verwendet. Offensichtlich kamen viele Fotos auf landwirtschaftlich kultivierten Flächen zu Stande. In der zu Beginn dieser Diskussion behandelten Dichtefunktion konnte festgestellt werden, dass A *near* B Beziehungen vorwiegend ausserhalb von Städten zu Stande kommen. Deshalb erstaunt es nicht, dass häufig Imageclasses verwendet werden, die charakteristisch für ländliche Bezeichnungen sind. Einige Imageclasses wie House, Railway, Railwaybridge usw. könnten durchaus in allen Gegenden verwendet werden. Weshalb jedoch genau Farmland so übermässig oft verwendet wurde, kann nicht erklärt werden.

Die Abbildung 5.20 reiht die Imageclasses der Häufigkeit nach auf. Um die Fragestellung zu beantworten, ob im Zusammenhang mit grösseren Imageclasses auch weitere Distanzen verwendet werden, entsteht das Problem der Klassifizierung der Imageclasses. Denn es ist nicht eindeutig, welche Imageclasses wie gross sind, somit ist eine Rangordnung nicht objektiv durchzuführen. Im Zusammenhang mit der Imageclass Moorland wurden im Mittel die mit Abstand grössten Distanzen verwendet. Obwohl subjektiv gesehen zum Beispiel die Imageclass Woodland genau

eine so grosse Ausdehnung haben kann, wie Moorland, liegt Woodland im Mittelfeld, das heisst mit Woodland werden im Schnitt ca. 1.5 km kürzere Distanzen verwendet als mit Moorland. Die gemäss GEO-DB kleinste Imageclass stellt Barn dar. Diese Klasse wird im Durchschnitt mit einer Distanz von etwa 1.2 km verwendet. Es kann festgestellt werden, dass die Wahl der Imageclass keinen Einfluss auf die Distanz hat.

6.5 Verteilung der Toponyme im Gazetteer

Es wurde untersucht, ob eine Verteilung, bei der jeder A-Koordinate aus der GEO-DB eine zufällig gewählte B-Koordinate von einem Toponym aus dem Gazetteer zugeordnet und die Distanz berechnet wurde (Gazetteer-Verteilung), über eine Ähnlichkeit verfügt zur Verteilung der Distanzen der GEO-DB. Dabei wurde darauf geachtet, dass es sich um denselben Ausschnitt handelt und dass gleich viele Stichproben ausgewertet werden. Wie erwartet verfügt die Gazetteer-Verteilung über einen massiv grösseren Mittelwert (im Norden rund 550 km, im Süden rund 450 km) als jene von GEOGRAPH (im Norden 2.46 km, im Süden 2.12 km). Dies war nicht anders zu erwarten, denn ansonsten würde dies bedeuten, dass die Stichproben der GEO-DB ebenfalls zufällig zu Stande gekommen wären.

Die Frage ist nun, ob die Gazetteer-Verteilung zur Verteilung der Stichproben von GEOGRAPH ähnlich ist und ob sie diese gar beeinflusst. Denn bei der Gazetteer-Verteilung ist der relative Nord-Süd-Unterschied des Mittelwertes ähnlich gross, wie der relative Nord-Süd-Unterschied der Distanzen von GEOGRAPH (-17.67% und -13.82%). Da auch die Verteilung der GEO-DB auf dem Gazetteer basiert (B-Koordinate), kann man die Vermutung anstellen, dass die Gazetteer-Verteilung den Unterschied zwischen Norden und Süden in der GEO-DB beeinflusst. Diese Annahme erscheint banal, wenn man vermutet, dass die Gazetteer-Verteilung bloss das Verhalten der Toponym-Dichte adaptiert. Betrachtet man nun jedoch den Vergleich zwischen städtischem und ländlichem Gebiet, stimmt der Zusammenhang zwischen der Toponym-Dichte und der verwendeten Distanz nicht. In der Stadt werden grössere Distanzen verwendet, obwohl eine höhere Dichte vorherrscht als auf dem Land. Die Gazetteer-Verteilung weist jedoch interessanterweise ein relativ ähnliches Verhalten auf zwischen städtischem und ländlichem Gebiet (-27.95% und -20.67%). Man kann also festhalten, dass die Gazetteer-Verteilung (zumindest in diesen Fällen) einen mächtigeren Einfluss auf die Stichproben hat als die eigentliche Toponym-Dichte in den entsprechenden Regionen.

Einzig der Vergleich zwischen dem Schottischen Hochland und dem Englischen Flachland liefert bei der Auswertung der Gazetteer-Verteilung und der Verteilung gemäss GEO-DB kein

so stark ähnliches Verhalten wie beim Stadt-Land- und Nord-Süd-Vergleich. Die Gazetteer-Verteilung und die Verteilung nach GEOGRAPH verfügen zwar beide über einen höheren Mittelwert im Flachland, jedoch ist die Differenz vom Flachland zum Hochland in der Gazetteer-Verteilung mit +3.17% nicht so gross wie die Differenz in der Verteilung von GEOGRAPH, die +20.74% beträgt. Weshalb in diesem Falle die Verteilung der Stichproben der GEO-DB ein anderes Verhalten zeigt als die Gazetteer-Verteilung, müsste nun mit Hilfe weiterer Faktoren wie Sichtbarkeit und Erreichbarkeit ergründet werden (Yao und Thill, 2005).

Kapitel 7

Schlussfolgerungen und Ausblick

7.1 Schlussfolgerungen

7.1.1 Durchschnittliche Verwendung

In sämtlichen Auswertungen kann grundsätzlich dasselbe Muster beobachtet werden: Mindestens 95% der Stichproben konzentrieren sich in einem Wertebereich zwischen 0 und 7 km. Im Mittel wird in ganz Grossbritannien im Zusammenhang mit der Kombination „*near* + Toponym“ eine Distanz von 2.28 km verwendet. Der Median (und ebenfalls die am häufigsten verwendete Distanz) von 1 km ist ebenfalls ein Indiz für die vorwiegend sehr kleinen Distanzen, in Anbetracht dessen, dass ein Stichproben-Spektrum von knapp 50 km untersucht wurde. Äusserst selten werden Distanzen von über 10 km verwendet (3.5%). Diese Stichproben sind möglicherweise durch Faktoren wie Bekanntheit, Attraktivität oder Erreichbarkeit der Toponyme zu Stande gekommen (Gahegan, 1995; Hernández et al., 1995; Guesgen, 1999; Yao und Thill, 2005).

7.1.2 Regionale Unterschiede

Im Norden Grossbritanniens werden im Vergleich zum Süden durchschnittlich grössere Distanzen verwendet. Dies kann darauf zurückzuführen sein, dass im Süden mit 7.3 Toponymen pro km^2 eine mehr als doppelt so hohe Toponym-Dichte herrscht als im Norden. Diese unterschiedliche Dichte führt dazu, dass die Verwendeung von *near* unterschiedliche Distanzen impliziert. Im Süden hat es innerhalb einer Distanz mehr potentielle Ziele, worauf sich der Teilnehmer beziehen kann. Dadurch ist die Wahrscheinlichkeit höher, dass naheliegende Toponyme gewählt werden.

Beim Stadt-Land-Vergleich ist ebenfalls die Toponym-Dichte dafür verantwortlich, dass in der Stadt Stichproben mit grösseren Distanzen vorhanden sind, obwohl paradoxerweise in der Stadt eine höhere Toponym-Dichte vorherrscht. Diese Dichte ist jedoch so hoch, dass ein „Umkehr-

Effekt“ eintritt: Aus kartographischen Gründen werden nicht alle Toponyme auf einer Karte eingetragen (Generalisierung), sprich: die kleineren Objekte müssen den grösseren bzw. wichtigeren weichen (Baumgartner, 1990). Dies führt dazu, dass mehr Beziehungen zu Stande kommen, die sich auf ein Toponym beziehen, das gross ist; grosse Toponyme liegen weiter auseinander als kleine. Somit entstehen in städtischen Gebieten die grösseren Distanzen als in ländlichen Regionen.

Bei der Analyse der durchschnittlich verwendeten Distanzen pro 100x100 km^2 Kachel in ganz Grossbritannien (siehe Abbildung 5.10, Seite 46) kann auf Grund des „Insel-Effektes“ ein Unterschied zwischen den Werten auf dem Festland und auf den Inseln bzw. Halbinseln festgestellt werden: Die Distanzen im Landesinnern sind im Mittel 1 km grösser als jene auf den Inseln. Dies ist darauf zurückzuführen, dass die Teilnehmer auf den Inseln sich auf einen kleineren Referenzrahmen beziehen, der durch den Umriss der Insel definiert wird. Auf dem Festland kann der Referenzrahmen theoretisch bis zur Grenze von Grossbritannien reichen. Ein grösserer Referenzrahmen beeinflusst die Wahrnehmung von der Distanz - man empfindet Distanzen als weniger weit, als in einem Gebiet, das einen kleinen Referenzrahmen hat (Gahegan, 1995).

7.1.3 Korrelationen

Es kann kein Zusammenhang gefunden werden zwischen der verwendeten Distanz im Zusammenhang mit *near* und der Einwohnerzahl des Toponyms. Dagegen kann beobachtet werden, dass Grossstädte konstant mit einer Distanz von mindestens 4 km erwähnt werden. Dies ist auf den sogenannten „Zentrums-Effekt“ zurückzuführen: Befindet sich der Teilnehmer in einer Stadt, bezeichnet er seinen Standort (logischerweise) nie damit, dass er sich *in der Nähe* der Stadt befindet (sondern *in* der Stadt). Befindet er sich 1 km vom Stadtrand entfernt, ist die Verwendung von *near* plus der Städtename wiederum angebracht. Da Städte im Gazetteer mit der Koordinate, die sich auf das Städtezentrum bezieht, attributiert sind, wird die Distanz zum Zentrum der Stadt berechnet (und nicht zum Stadtrand). Im Zusammenhang mit einer Grossstadt weist die Beziehungen A *near* B nie eine kleinere Distanz als 4 km auf.

Es ist kein Zusammenhang zwischen den angegebenen Imageclasses und der Distanz zwischen A und B zu erkennen. Es kann keine Struktur erkennt werden, ebenfalls kann nicht begründet werden, weshalb mit der Imageclass „Barn“ die kleinsten und mit „Moorland“ mit Abstand die grössten Distanzen verwendet werden.

7.1.4 Verteilung der Toponyme im Gazetteer

Ein weiterer Grund, weshalb regionale Unterschiede entstehen, kann auf die zufällige Verteilung der Toponyme im Gazetteer zurückgeführt werden: Es wurde eine zufällige Verteilung generiert, in dem man den A-Koordinaten aus der GEO-DB zufällige B-Koordinaten (desselben Gebietes) aus dem Gazetteer zuordnete. Dabei können zwischen zwei Untersuchungsgebieten in beiden Verteilungen dieselben relativen Differenzen der entstanden Distanzen beobachtet werden. Damit ist dies neben der Toponym-Dichte und dem Referenzrahmen die dritte Erklärung für regional unterschiedliche Distanzen, die im Rahmen dieser Arbeit ergründet wurden.

7.1.5 Granularität der Daten

Die Koordinaten von A und B liegen in der GEO-DB und im Gazetteer mit einer Auflösung von 1 km^2 vor. Da sich über 50% aller Stichproben zwischen 0 und 1 km abspielen, könnte anhand einer höheren Granularität (z.B. von 0.1 km^2) ein feineres Verhalten der Beziehung A *near* B herauskristallisiert werden. Dies würde noch differenziertere Vergleiche zwischen verschiedenen Regionen erlauben.

7.2 Ausblick

Eingangs dieser Arbeit wurde erläutert, dass das kognitive Basiswissen der Teilnehmer erörtert wird, um dieses in GI-Systemen implementieren zu können. Diesbezüglich ist es denkbar, die Resultate zum Beispiel in Form eines „*near*-Tools“ in einem GI-Programm zu verwenden. Dieses Tool wäre demnach im Stande (ausgehend von einer Koordinate A) einen Sektor zu berechnen, innerhalb diesem sich der Mensch von jedem Punkt aus nahe zur Koordinate A fühlt.

Interessant wären weitere Forschungen über die intuitive Anwendung von *near* in Grossbritanninen bei der man sich auf Informationen stützen kann, die über eine höhere Auflösung als 1 km^2 verfügen. Es wäre jedoch auch spannend mit derselben Granularität wie in dieser Arbeit Untersuchungen in anderen Ländern, die flächenmässig Grossbritannien überlegen sind, anzustellen (vgl. 6.2.5). Es ist durchaus vorstellbar, dass dann eine Auflösung von 1 km^2 reichen könnte, um differenziertere Muster zu beobachten.

Weiterführende Untersuchungen über die Beeinflussung der Verteilung der Toponyme im Gazetteer wären ebenfalls interessant. In der Literatur wurde bis jetzt ein möglicher Zusammenhang zwischen dem Gazetteer und den Stichproben nicht diskutiert. Falls auch weitere Untersuchungen diesen Zusammenhang erhärten können, würde dies bedeuten, dass künftig Forschungen über

räumliche Beziehungen in Grossbritannien, die sich auf das 1:50'000 Gazetteer stützen, dieses Verhalten berücksichtigen müssten.

Diese Arbeit beschäftigt sich mit dem räumlichen Relationsbegriff *near*. Mit derselben Datenbank könnte man durchaus auch Untersuchungen durchführen, die sich mit weiteren Begriffen wie zum Beispiel *right*, *left*, *north of*, *south of*, usw. beschäftigen. Danach könnten die Resultate in einem Katalog abgespeichert werden, der für jede Region in Grossbritannien für jeden räumlichen Begriff einen Sektor definiert. Auf diesen Katalog könnte nun eine Internetsuchmaschine bei einer entsprechenden Abfrage zugreifen. Somit wüsste die Suchmaschine in welchem Sektor sie nach einem Artikel suchen muss, um dem Benutzer adäquate Resultate zu liefern.

Literaturverzeichnis

Baumgartner, U. (1990): Kartographisches Generalisieren, Kapitel Generalisierung topographischer Karten. Schweizerische Gesellschaft für Kartographie, Zürich, S. 23–24.

Belnap, N. (1977): Modern Uses of Multiple-Valued Logic, Kapitel A useful four-valued logic. Dordrecht-Boston: Reidel, S. 5–37.

Bill, R. (1999): Grundlagen der Geo-Informationssysteme - Analysen, Anwendungen und neue Entwicklungen, 2. Auflage. Herbert Wichmann, Heidelberg.

Bonini, N., Osherson, D., Viale, R. und Williamson, T. (1999): On the Psychology of Vague Predicates. In: Mind & Language, Blackwell Publishers, 14(4), S. 377–393.

Brennan, J. und Martin, E. (2006): Membership Functions for Spatial Proximity. In: Lecture Notes in Computer Science, 4304, S. 942–949.

Ebdon, D. (1977): Statistics in Geography - A Practical Approach. Basil Blackwell, Oxford.

Edwardes, A. J., Purves, R. S., Bircher, S. und Matyas, C. (2008): Concept ontology experimental report. In: Tripod (EC IST 6th Framework Programme Project No. 045335) - D.1.4.

Egenhofer, M. J. und Mark, D. M. (1995): Naive Geography. In: Spatial Information Theory. Departament of Computer Science, University of Maine, S. 1–15.

Fabrikant, S. I. und Buttenfield, B. P. (2001): Formalizing Semantic Spaces for Information Access. In: Annals of the Association of American Geographers, 91, S. 263–280.

Fisher, P. (2000): Sorites paradox and vague geographies. In: Fuzzy Sets and Systems, 113, S. 7–18.

Frank, A. U. (1996): Qualitative Spatial Reasoning: Cardinal Directions as an Example. In: International Journal of Geographical Information Science, 10(3), S. 269–290.

Friedmann, A. und Brown, N. (2000): Reasoning about geography. In: Journal of Experimental Psychology, 129(2), S. 193–219.

Gahegan, M. (1995): Proximity operators for qualitative spatial reasoning. In: Lecture Notes in Computer Science, 988, S. 31–44.

Guesgen, H. W. (1999): Reasoning with Words about Geographic Information. In: IJCAI '97: Selected and Invited Papers from the Workshop on Fuzzy Logic in Artificial Intelligence. Springer-Verlag, London, UK, S. 133–148.

Guesgen, H. W. (2002): Reasoning About Distance Based on Fuzzy Sets. In: Applied Intelligence, 17(3), S. 265–270.

Hard, G. und Gliedner, A. (1979): Die Wahre Landschaft. Eine kritische Analyse des Landschaftsbegriffs, Kapitel Wort und Begriff Landschaft anno 1976. Der Neue Brockhaus. V, 325-326., S. 16–24.

Hernández, D. (1994): Qualitative Representation of Spatial Knowledge, Band 804 von Lecture Notes in Computer Science. Springer.

Hernández, D., Clementini, E. und Felice, P. D. (1995): Qualitative Distances. In: Spatial Information Theory: a theoretical basis for GIS, 988, S. 45–58.

Jones, C. B., Purves, R. S., Clough, P. D. und Joho, H. (2008): Modelling vague places with knowledge from the Web. In: International Journal of Geographical Information Science, S. 1365–8816.

Kuipers, B. und Levit, T. (1990): Advances in Spatial Reasoning, Kapitel Navigation and Mapping in Large-Scale Space. S. Chen, Ablex Publishing Corp., S. 207 – 251.

Lang, B. (2005): Vorlesung Bildverarbeitung, Fachhochschule Osnabrück. In: Vorlesungsunterlagen von Prof. Dr.-Ing. Bernhard Lang, S. 27.

Lichtner, W. (1981): Wissenschaftliche Arbeiten der Fachrichtung Vermessungswesen der Universität Hannover, Kapitel Anwendungsmöglichkeiten der Rasterdatenverarbeitung in der Kartographie. Universität Hannover, S. 79.

Loerch, U. und Guesgen, H. W. (1997): Qualitative spatial reasoning under uncertainity in geographical information system. IEEE.

McNamara, T. P. (1991): Memory's view of space. In: The psychology of learning and motivation: Advances in research and theory, 27, S. 147–186.

Montello, D. (1998): Spatial and temporal reasoning in geographic information systems. Oxford University Press, Oxford.

Montello, D. R. (1993): Spatial Information Theory A Theoretical Basis for GIS, Band 716 von Lecture Notes in Computer Science, Kapitel Scale and multiple psychologies of space. Springer, Berlin, S. 312–321.

Orgin, D. (1999): Landscape as a research problem. In: Agriculturale Conspectus Scientificus, 64(4).

Pullar, D. und Egenhofer, M. (1988): Toward Formal Definitions of Topological Relations among Spatial Objects. In: 3rd International Symposium on Spatial Data Handling. International symposium on SDH, S. 21–29.

Purves, R. S. und Edwardes, A. J. (2008): Exploiting Volunteered Geographic Information to describe Place. In: GIS Research UK (GISRUK 2008), Manchester Metropolitan University.

Purves, R. S., Edwardes, A. J. und Sanderson, M. (2008): Describing the Where - improving image annotation and search through geography. In: First Intl. Workshop on Metadata Mining for Image Understanding (MMIU 2008), Madeira, Portugal.

Robinson, B. (1990): Interactive machine acquisition of a fuzzy spatial relation. In: Computers & Geosciences, Artificial intelligence applications in geoscience, 16(6), S. 857–872.

Sadalla, E. K., Burroughs, W. J. und Staplin, L. J. (1980): Reference points in spatial cognition. In: J. exp. Psychol.: hum. Learn.Mem, 6(5), S. 516–528.

Sainsbury, R. M. (1995): Paradoxes. Cambridge University Press.

Samet, H. (1990): The Design and Analysis of Spatial Data Structures. Addison-Wesley series in computer science.

Stevens, A. und Coupe, P. (1978): Distortions in Judged Spatial Relations. In: Cognitive Psychology, 10(4), S. 422–437.

Tang, L. (1991): Einsatz der Rasterdatenverarbeitung zum Aufbau digitaler Geländemodelle, Band 73. Geodätische Institute der Technischen Universität Graz.

Worboys, M. (1996): Metrics and Topologies for Geographic Space. In: Advances in Geographic Information Systems Research II: Proceedings of the International Symposium on Spatial Data Handling, Delft. International Geographical Union, S. 7A.1–7A.11.

Worboys, M. (2001): Nearness relations in environmental space. In: International Journal of Geographical Information Science, 15, S. 633–651.

Worboys, M., Duckham, M. und Kulik, L. (2004): Commonsense notions of proximity and direction in environmental space. In: Spatial Cognition and Computation, 4(4), S. 285–312.

Yao, X. und Thill, J.-C. (2005): How Far Is Too Far? A Statistical Approach to Context-contingent Proximity Modeling. In: Transactions in GIS, 9(2), S. 157–178.

Yao, X. und Thill, J.-C. (2006): Spatial queries with qualitative locations in spatial information systems. In: Computers, Environment and Urban Systems, 30(4), S. 485–502.

Yao, X. und Thill, J. C. (2007): Neurofuzzy Modeling of Context - Contingent Proximity Relations. In: Geographical Analysis, 39, S. 169–194.

Anhang A

Beispiel zweier Perl-Skripts

„A *near* B“ - Extraktion


```
#!/usr/bin/perl
#use strict;
#use warnings;

open (ONLYNEAR, '<', 'only_near_clean.txt');          #öffnet Datei onlynear.txt ohne zu überschreiben (<),
open (POPTAB, '<', 'clean_pop_gb.txt');               #muss im gleichen directory sein. Einwohnertabelle
                                                      #von GB (bereinigt, siehe perlprogramm
                                                      #poptab_clea_only_england) wird eingelesen
open (GAZ, '<', '50kgaz2006.txt');


while (<ONLYNEAR>) {                                  #while <> : zeilenweise wird folgendes Prg ausgeführt

   my @line_splitted = split(/\t/);                   #jede Zeile wird von Tab zu Tab in einzelne Strings geteilt und
                                                      #ins Array @line_splited gespeichert
    $line_splitted[1] =~ s/Near/near/;                #ersetzt in titel und kommentar ein grossgeschriebenes Near zu
                                                      #einem kleinen near. Somit findet es unten mit der funktion
                                                      #INDEX das near.
    $line_splitted[2] =~ s/Near/near/;
```

```
    ####TOPONYM B extrahieren####
    my $topoB_in_title = generate_clean_toponymB($line_splitted[1]);          #subroutine generate_clean_toponymB wird
                                                                              #mit TITLE gefüttert

    my $topoB_in_comment = generate_clean_toponymB($line_splitted[2]);        #wenn near im Titel und Kommentar
                                                                              #vorkommt, welches ist
                                                                              #dann Topo B? Es zeigt beide
                                                                              #B's aber mit dem gleichen OSGB A.

    ####OSGB A extrahieren####
    my $koordA = $line_splitted[8];

    ####OSGB B extrahieren aus Gazetteer####
    my $koordB_title = get_koordB_title($topoB_in_title);
    my $koordB_comment = get_koordB_comment($topoB_in_comment);

    ####EINWOHNERZAHL B extrahieren aus Populationstabelle GB####
    my $populationB_in_title = get_pop_from_topoB_in_title($topoB_in_title); #der Subroutine das Topo vom Titel
                                                                              #übergeben
    $populationB_in_title =~ s/\n|\r//g;                                      #enfernt <cr> (carriage return),
                                                                              #ohne diesem RA (regulären Ausdruck),
                                                                              #würde im output immer ein newline
                                                                              #eingefügt nach der Einwohnerzahl

    my $populationB_in_comment =
                            get_pop_from_topoB_in_comment($topoB_in_comment);#der Subroutine das Topo vom
                                                                              #Kommentar übergeben
    $populationB_in_comment =~ s/\n|\r//g;                                    #enfernt <cr> (carriage return), ohne
                                                                              #diesem RA (regulären ausdruck), würde
                                                                              #im output immer ein newline eingefügt
                                                                              #nach der Einwohnerzahl

    ####Grid_ID, Pers_ID und Imageclass von Geograph####
    my $grid_id = $line_splitted[0];
    my $pers_id = $line_splitted[9];
    my $imageclass = $line_splitted[3];

    print "$grid_id\t";
```

```
print "$topoB_in_title\t";
print "$topoB_in_comment\t";
print "$koordA\t";
print "$koordB_title\t";
print "$koordB_comment\t";
print "$populationB_in_title\t";
print "$populationB_in_comment\t";
print "$imageclass\t";
print "$pers_id\n";
}


close (POPTAB);
close (ONLYNEAR);
close (GAZ);

sub generate_clean_toponymB{

my $toponymB  = "";
my $topo = $_[0];                                   #$_[0] ist der TITLE bzw. der COMMENT
my $near = "near";

if (check_near($topo)) {
        my $near_position = index($topo, $near) + 1;
            if ($near_position == 0){               #wenn es kein "near" im String gibt (gleich 0), dann
                                                    #verlässt das Prg die Subroutine,
                return;
                }

        my $near_right = substr($topo, $near_position);     #ab n-ten Buchstabe von $topo (n = $near_position)
                                                            #wird neuer String generiert. Speichert dies in
        my @right = split(/\s/, $near_right);               #$near_right splited $near_right nach white-spaces
                                                            #auf und speichert array in @right
        my $i = 1;

        while (check_case($right[$i])) {
            $toponymB = $toponymB." ".$right[$i];           #jedes Wort mit Grossbuchst wird in $toponym gespeichert.
                                                            #bei jedem Durchgang wird $toponym neu abgefüllt mit
            $i++;                                           #zusätzlichen Wörtern, die Grossbuchst am Anfang haben
```

```
            }

            $toponymB =~ s/\.//g;                    #s/Suchmuster/Ersatzzeichenkette/g  g ersetzt ALLE Punkte
                .

            return $toponymB;                        #hier werden alle Punkte entfernt.
        }                                            #gibt string $toponymB an prg zurück
}


sub check_near{
        return (/\bnear\b/i);                        #sub schaut, ob in einem Satz "near" alleinstehend
                                                     #(-> \b...\b), (gross/kleinschreibung egal -> i) vorkommt
}


sub check_case{                                      #sub schaut, ob erstes Wort nach near ($right[1]) mit
                                                     #Grossbuchstabe beginnt
        return ($_[0] =~ /[A-Z]/);                   #Mustererkennung schaut ob ein Teil des Strings identisch
                                                     #ist, bei Wort, das gross beginnt ist es das
}


sub get_koordB_title {
   seek(GAZ,0,0);

   while (<GAZ>) {                                   #while <> : zeilenweise wird folgendes Prg ausgeführt
      my @gaz_line_splitted = split(/:/);

      $_[0]=~ s/^\s+//;                              #entfernt Leerschlag vor Topo in Titel bzw Comment.
                                                     #Ohne dies wäre es nie ($_[0] eq $gaz_line_splitted[2]),
                                                     #da vor $_[0] immer ein Leerschlag wäre und somit nie
                                                     #equal wäre.

       if ($_[0] eq $gaz_line_splitted[2]) {
          return $gaz_line_splitted[1];
          }
   }
```

```
}


sub get_koordB_comment{
   seek(GAZ,0,0);

   while (<GAZ>) {                                  #while <> : zeilenweise wird folgendes Prg ausgeführt
      my @gaz_line_splitted = split(/:/);

      $_[0]=~ s/^\s+//;                             #entfernt Leerschlag vor Topo in Titel bzw Comment.
                                                    #ohne dies wäre es nie ($_[0] eq $gaz_line_splitted[2]),
                                                    #da vor $_[0] immer ein Leerschlag wäre und somit nie
                                                    #equal wäre.

      if ($_[0] eq $gaz_line_splitted[2]) {
         return $gaz_line_splitted[1];
         }
   }
}


sub get_pop_from_topoB_in_title {
   seek(POPTAB,0,0);

   while (<POPTAB>) {
      my @poptab_line_splitted = split(/\t/);

      if ($_[0] eq $poptab_line_splitted[0]) {      #$_[0] ist übergebener Wert ($topoB_in_title, sprich
                                                    #das Toponym B). $poptab_line_splitted[0] ist Toponym
                                                    #in der Einwohnertabelle $poptab_line_splitted[1]
         return $poptab_line_splitted[1];           #ist die Einwohenrzahl aus POPTAB
      }
   }
}


sub get_pop_from_topoB_in_comment {
   seek(POPTAB,0,0);
```

```
while (<POPTAB>) {
    my @poptab_line_splitted = split(/\t/);

    if ($_[0] eq $poptab_line_splitted[0]) {

        return ($poptab_line_splitted[1]);
    }
}
}
```

```
#$_[0] ist übergebener Wert ($topoB_in_title, sprich
#das Toponym B). $poptab_line_splitted[0] ist Toponym
#in der Einwohnertabelle $poptab_line_splitted[1]
#ist die Einwohenrzahl aus POPTAB
```

Distanzberechnung

```
#!/usr/local/bin/perl
#use strict;
#use warnings;

open (GEOGRAPH, "<", "combi_def.txt");

while (<GEOGRAPH>) {
   my @geogr_array = split(/\t/);

      print "Grid_ID: $geogr_array[0]\n";
      print "OSGB_A: $geogr_array[3]\n";

         $koord_A = get_buchstaben($geogr_array[3]);
         print "KOORD_A X:Y: $koord_A\n";



   ###BERECHNUNG VON OSGB B TITLE###
   if($geogr_array[4] eq ""){
      print "Kein OSGB B-Title vorhanden\n";
         goto COMMENT;
   }
            else{
               print "OSGB_B_Title: $geogr_array[4]\n";
               $koord_B_Title = get_buchstaben($geogr_array[4]);
                  print "KOORD_B_Title X:Y: $koord_B_Title\n";

                     my $distance_A_B_Title = get_distance($koord_A, $koord_B_Title
                          );
                        print "Distanz zw A und B-Title ist $distance_A_B_Title km\
                            n";
            }


   ###BERECHNUNG VON OSGB B COMMENT###
   COMMENT:
   if($geogr_array[5] eq ""){
      print "Kein OSGB-Comment vorhanden\n";
   }
            else{
               print "OSGB_B_Comment: $geogr_array[5]\n";
               $koord_B_Comment = get_buchstaben($geogr_array[5]);
                 print "KOORD_B_Comment X:Y: $koord_B_Comment\n";

                      my $distance_A_B_Comment = get_distance($koord_A,
                           $koord_B_Comment);
                        print "Distanz zw A und B-Comment ist $distance_A_B_Comment
                            km\n";
            }
```

```
    END:

    print "\n";
}
close (GEOGRAPH);


sub get_buchstaben{

    my @ersetzung = split(//, $_[0]);

        print "ERSTER BUCHSTABE $ersetzung[0]\n";
        print "ZWEITER BUCHSTABE $ersetzung[1]\n";

    my $buchstaben = "$ersetzung[0]"."$ersetzung[1]";
        print "Nur Buchstaben: $buchstaben\n";

    ###UMWANDLUNG BUCHSTABEN IN ZAHL (km)###
    $buchstaben =~ s/HP/400\t1200\t/;
    $buchstaben =~ s/HT/300\t1100\t/;
    $buchstaben =~ s/HU/400\t1100\t/;
    .           .           .
    .           .           .          # Konvertierung der Kombinationen HW bis SW
    .           .           .
    $buchstaben =~ s/SX/200\t000\t/;
    $buchstaben =~ s/SY/300\t000\t/;
    $buchstaben =~ s/SZ/400\t000\t/;
    $buchstaben =~ s/TV/500\t000\t/;

    @splitted_umwandlung = split(/\t/, $buchstaben);

    my $x_koord = $splitted_umwandlung[0]+($ersetzung[2].$ersetzung[3]);
    my $y_koord = $splitted_umwandlung[1]+($ersetzung[4].$ersetzung[5]);

           print "X KORDINATE AUFGETEILT:"."$splitted_umwandlung[0]"."+"."$ersetzung
               [2]"."$ersetzung[3]\n";
           print "Y KORDINATE AUFGETEILT:"."$splitted_umwandlung[1]"."+"."$ersetzung
               [4]"."$ersetzung[5]\n";

    return $x_koord.":".$y_koord;
}


sub get_distance{

    my @splitted_koord_A = split(/:/, $_[0]);
    my @splitted_koord_B = split(/:/, $_[1]);

    $kath_a = $splitted_koord_A[0]-$splitted_koord_B[0];
    $kath_b = $splitted_koord_A[1]-$splitted_koord_B[1];
```

```
    $hypo_c = sqrt(($kath_a)**2 + ($kath_b)**2);

    $distanz = sprintf("%.2f",$hypo_c);

    return $distanz;
}
```

Anhang B

Häufigkeitsverteilung aller gültigen Stichproben

Distanz (km)	Häufigkeit	Prozent	Kumulierte Prozente
0.00	2963	19.94	19.94
1.00	5012	33.73	53.66
1.41	2425	16.32	69.98
2.00	865	5.82	75.80
2.24	1160	7.81	83.61
2.83	246	1.66	85.26
3.00	183	1.23	86.49
3.16	310	2.09	88.58
3.61	247	1.66	90.24
4.00	58	0.39	90.63
4.12	105	0.71	91.34
4.24	63	0.42	91.76
4.47	105	0.71	92.47
5.00	91	0.61	93.08
5.10	40	0.27	93.35
5.39	42	0.28	93.63
5.66	14	0.09	93.73
5.83	55	0.37	94.10
6.00	11	0.07	94.17
6.08	32	0.22	94.39
6.32	33	0.22	94.61
6.40	32	0.22	94.83
6.71	27	0.18	95.01
7.00	9	0.06	95.07
7.07	17	0.11	95.18
7.21	16	0.11	95.29
7.28	13	0.09	95.38
7.62	13	0.09	95.46
7.81	17	0.11	95.58
8.00	8	0.05	95.63

8.06	40	0.27	95.90
8.25	6	0.04	95.94
8.49	4	0.03	95.97
8.54	9	0.06	96.03
8.60	16	0.11	96.14
8.94	15	0.10	96.24
9.00	5	0.03	96.27
9.06	10	0.07	96.34
9.22	9	0.06	96.40
9.43	4	0.03	96.43
9.49	4	0.03	96.45
9.85	6	0.04	96.49
9.90	4	0.03	96.52
10.00	3	0.02	96.54
10.05	2	0.01	96.55
10.20	2	0.01	96.57
10.30	5	0.03	96.60
10.44	3	0.02	96.62
10.63	8	0.05	96.68
10.77	1	0.01	96.68
10.82	5	0.03	96.72
11.00	1	0.01	96.72
11.05	1	0.01	96.73
11.18	10	0.07	96.80
11.31	1	0.01	96.80
11.40	7	0.05	96.85
11.66	7	0.05	96.90
11.70	4	0.03	96.92
12.04	7	0.05	96.97
12.08	4	0.03	97.00
12.17	3	0.02	97.02
12.21	3	0.02	97.04
12.37	1	0.01	97.05
12.53	1	0.01	97.05
12.65	3	0.02	97.07
12.73	2	0.01	97.09
12.81	3	0.02	97.11
13.00	2	0.01	97.12
13.04	5	0.03	97.15
13.15	3	0.02	97.17
13.34	6	0.04	97.21
13.42	2	0.01	97.23
13.45	5	0.03	97.26
13.60	6	0.04	97.30
13.89	3	0.02	97.32
14.00	2	0.01	97.34
14.04	1	0.01	97.34
14.14	5	0.03	97.38
14.32	6	0.04	97.42
14.42	2	0.01	97.43
14.76	2	0.01	97.44

14.87	5	0.03	97.48
15.00	2	0.01	97.49
15.03	2	0.01	97.50
15.13	5	0.03	97.54
15.23	1	0.01	97.54
15.26	7	0.05	97.59
15.30	2	0.01	97.60
15.52	1	0.01	97.61
15.56	2	0.01	97.62
15.62	1	0.01	97.63
15.65	2	0.01	97.64
15.81	4	0.03	97.67
16.03	3	0.02	97.69
16.12	4	0.03	97.72
16.16	2	0.01	97.73
16.28	2	0.01	97.75
16.40	1	0.01	97.75
16.55	1	0.01	97.76
16.76	1	0.01	97.77
17.20	3	0.02	97.79
17.26	2	0.01	97.80
17.49	3	0.02	97.82
17.69	1	0.01	97.83
17.72	2	0.01	97.84
17.89	1	0.01	97.85
18.03	2	0.01	97.86
18.11	1	0.01	97.87
18.25	3	0.02	97.89
18.36	1	0.01	97.89
18.38	3	0.02	97.91
18.44	2	0.01	97.93
18.68	1	0.01	97.93
18.79	3	0.02	97.95
18.97	1	0.01	97.96
19.00	2	0.01	97.97
19.03	2	0.01	97.99
19.21	3	0.02	98.01
19.31	4	0.03	98.04
19.42	1	0.01	98.04
19.70	3	0.02	98.06
19.80	1	0.01	98.07
19.85	1	0.01	98.08
20.10	1	0.01	98.08
20.22	2	0.01	98.10
20.25	1	0.01	98.10
20.62	1	0.01	98.11
20.88	1	0.01	98.12
21.02	1	0.01	98.12
21.10	2	0.01	98.14
21.19	1	0.01	98.14
21.21	1	0.01	98.15

21.38	1	0.01	98.16
21.54	1	0.01	98.16
21.59	2	0.01	98.18
21.84	2	0.01	98.19
21.93	3	0.02	98.21
22.20	2	0.01	98.22
22.36	3	0.02	98.24
22.47	1	0.01	98.25
22.56	1	0.01	98.26
22.67	2	0.01	98.27
22.80	1	0.01	98.28
22.85	3	0.02	98.30
23.02	1	0.01	98.30
23.09	2	0.01	98.32
23.26	1	0.01	98.32
23.32	1	0.01	98.33
23.35	2	0.01	98.34
23.41	2	0.01	98.36
23.54	2	0.01	98.37
23.77	1	0.01	98.38
23.85	1	0.01	98.39
24.02	1	0.01	98.39
24.17	1	0.01	98.40
24.21	1	0.01	98.41
24.33	1	0.01	98.41
24.41	1	0.01	98.42
24.52	1	0.01	98.43
24.70	1	0.01	98.43
25.06	1	0.01	98.44
25.08	1	0.01	98.45
25.18	1	0.01	98.45
25.46	1	0.01	98.46
25.50	1	0.01	98.47
25.55	1	0.01	98.47
25.61	2	0.01	98.49
25.63	1	0.01	98.49
25.71	1	0.01	98.50
25.94	2	0.01	98.51
26.08	2	0.01	98.53
26.17	1	0.01	98.53
26.31	1	0.01	98.54
26.40	2	0.01	98.55
26.68	2	0.01	98.57
27.00	3	0.02	98.59
27.02	1	0.01	98.59
27.17	2	0.01	98.61
27.20	1	0.01	98.61
27.46	1	0.01	98.62
27.51	1	0.01	98.63
27.66	6	0.04	98.67
27.73	2	0.01	98.68

27.78	1	0.01	98.69
27.80	1	0.01	98.69
27.89	4	0.03	98.72
28.02	3	0.02	98.74
28.18	1	0.01	98.75
28.23	1	0.01	98.76
28.32	1	0.01	98.76
28.43	1	0.01	98.77
28.44	1	0.01	98.78
28.64	2	0.01	98.79
28.86	1	0.01	98.80
29.00	2	0.01	98.81
29.07	2	0.01	98.82
29.12	1	0.01	98.83
29.15	2	0.01	98.84
29.41	1	0.01	98.85
29.53	1	0.01	98.86
29.55	1	0.01	98.86
29.73	2	0.01	98.88
30.02	2	0.01	98.89
30.07	2	0.01	98.90
30.36	1	0.01	98.91
30.68	2	0.01	98.92
30.81	2	0.01	98.94
31.00	2	0.01	98.95
31.02	2	0.01	98.96
31.05	1	0.01	98.97
31.40	1	0.01	98.98
31.58	1	0.01	98.98
31.83	2	0.01	99.00
32.00	2	0.01	99.01
32.02	4	0.03	99.04
32.06	1	0.01	99.04
32.14	1	0.01	99.05
32.25	3	0.02	99.07
32.39	1	0.01	99.08
32.56	1	0.01	99.08
32.57	2	0.01	99.10
32.65	1	0.01	99.11
32.76	2	0.01	99.12
33.02	4	0.03	99.15
33.06	1	0.01	99.15
33.24	4	0.03	99.18
33.30	1	0.01	99.19
33.53	1	0.01	99.19
33.60	1	0.01	99.20
33.62	1	0.01	99.21
33.84	2	0.01	99.22
33.97	1	0.01	99.23
34.01	1	0.01	99.23
34.06	1	0.01	99.24

34.13	2	0.01	99.25
34.21	1	0.01	99.26
34.48	5	0.03	99.29
34.67	1	0.01	99.30
34.83	1	0.01	99.31
34.93	1	0.01	99.31
34.99	2	0.01	99.33
35.00	1	0.01	99.33
35.23	1	0.01	99.34
35.34	1	0.01	99.35
35.47	2	0.01	99.36
35.74	1	0.01	99.37
35.78	2	0.01	99.38
36.06	1	0.01	99.39
36.24	1	0.01	99.39
36.50	1	0.01	99.40
36.67	1	0.01	99.41
36.88	1	0.01	99.41
36.89	1	0.01	99.42
37.20	3	0.02	99.44
37.59	1	0.01	99.45
37.64	1	0.01	99.45
38.08	1	0.01	99.46
38.28	1	0.01	99.47
38.29	2	0.01	99.48
38.33	2	0.01	99.50
38.60	2	0.01	99.51
38.83	1	0.01	99.52
38.90	1	0.01	99.52
39.05	4	0.03	99.55
39.12	1	0.01	99.56
39.20	2	0.01	99.57
39.56	1	0.01	99.58
39.81	2	0.01	99.59
39.82	1	0.01	99.60
39.85	1	0.01	99.60
39.96	1	0.01	99.61
40.00	1	0.01	99.62
40.02	1	0.01	99.62
40.05	1	0.01	99.63
40.31	2	0.01	99.64
40.61	1	0.01	99.65
41.05	1	0.01	99.66
41.11	1	0.01	99.66
41.23	2	0.01	99.68
41.34	1	0.01	99.68
41.40	1	0.01	99.69
41.59	2	0.01	99.70
41.62	1	0.01	99.71
42.01	2	0.01	99.72
42.05	1	0.01	99.73

42.20	1	0.01	99.74
42.45	1	0.01	99.74
42.52	1	0.01	99.75
42.95	1	0.01	99.76
43.00	2	0.01	99.77
43.01	1	0.01	99.78
43.28	1	0.01	99.78
43.57	1	0.01	99.79
43.86	1	0.01	99.80
43.93	1	0.01	99.80
44.15	2	0.01	99.82
44.18	1	0.01	99.83
44.27	2	0.01	99.84
44.42	1	0.01	99.85
44.69	2	0.01	99.86
44.78	1	0.01	99.87
44.91	1	0.01	99.87
45.00	1	0.01	99.88
45.12	1	0.01	99.89
45.88	1	0.01	99.89
46.57	1	0.01	99.90
46.82	1	0.01	99.91
47.01	1	0.01	99.91
47.07	4	0.03	99.94
47.63	2	0.01	99.95
47.80	1	0.01	99.96
48.00	1	0.01	99.97
48.10	1	0.01	99.97
48.17	1	0.01	99.98
48.88	2	0.01	99.99
49.48	1	0.01	100.00
Gesamt	14861	100.00	

Tabelle B.1: Häufigkeitstabelle aller 14'861 gültigen Stichproben von GEOGRAPH.