Landschaftsschönheit in Texten

Modellierung der Landschaftsqualität mit Opinion Mining

Masterarbeit GEO 511

Zürich, 23. August 2013

Vorgelegt von

Mario Nowak
Schwamendingenstr. 86
8050 Zürich
mario.nowak@uzh.ch
079 681 20 48
Matrikelnummer: 06-919-302

Betreut von

Prof. Dr. Ross Purves
ross.purves@geo.uzh.ch

Fakultätsvertretung
Prof. Dr. Ross Purves

Abteilung Geocomputation
Geographisches Institut – Universität Zürich

Zusammenfassung

Begriffe wie «Zersiedelung» und «haushälterische Bodennutzung», deren häufige Verwendung auffällt, und die in jüngster Zeit in der Schweiz durchgeführten Volksabstimmungen zum Raumplanungsgesetz und zur Zweitwohnungsinitiative zeugen davon, dass natürliche Landschaften als etwas Schützenswertes und gleichzeitig auch Bedrohtes wahrgenommen werden. Aber wo befinden sich besonders schützenswerte Landschaften und wie sieht das räumliche Muster der wahrgenommenen Landschaftsqualität aus?

Diese Arbeit untersucht eine neue Methode zur Beantwortung solcher Fragen. Die meisten bisherigen Untersuchungen zur Wahrnehmung von Landschaftsqualität basieren auf Probandenbefragungen mit beschränkter räumlicher Ausdehnung. Neue Datensätze im Internet erlauben aber neue Ansätze. In diesem Werk werden zwei britische Datensätze verwendet – der eine hat das Ziel, jede Quadratkilometerzelle Grossbritanniens mit einer Fotografie und nach Möglichkeit mit einem kurzen Text zu dokumentieren. Der andere enthält Landschaftsbewertungen von Freiwilligen, die auf einer Auswahl der Bilder des erstgenannten Datensatzes basieren. Das Ziel der Arbeit besteht darin, aus der Kombination dieser beiden Datensätze die Schönheit von Landschaften auf der Basis von Texten zu schätzen. Die Motivation ist, dass durch solche Methoden neue Datenquellen zur Evaluierung von Landschaften erschlossen werden können, die bisher nicht dazu verwendet wurden.

In einem ersten Schritt wurden die Muster in den Landschaftsbeschreibungen aufgedeckt. Dabei zeigte sich, dass Begriffe des natürlichen Raums besser bewertet sind als solche des urbanen Raums. Das führte zur Erkenntnis, dass zur Beantwortung der zweiten Forschungsfrage, nämlich der Modellierung der Landschaftsbewertungen aufgrund von Texten, ein generisches Stimmungslexikon ungeeignet ist. Stattdessen wurden landschaftsspezifische Schönheitslexika erstellt und ihre Leistungsfähigkeit untersucht.

Es wurden verschiedene Methoden zur Schätzung der Landschaftsbewertungen implementiert. Die besten Resultate wurden mit einer Einzelwortbetrachtung (Unigramme) der Beschreibungstexte erzielt. Damit konnte bei Einschränkung der maximalen Stimmvarianz 53.2 % der beobachteten Bewertungsvarianz erklärt werden, die Rangreihenfolge der Bewertungen konnte mit einer Genauigkeit von 64.8 % reproduziert werden. Bei unbeschränkter Stimmvarianz konnte 39.1 % der Bewertungsvarianz erklärt werden. Der landschaftliche Wert von Bildern, bei deren Bewertung Einigkeit geherrscht hat, konnte aufgrund der Texte also präziser geschätzt werden.

Der Beitrag dieser Arbeit ist, dass gezeigt wurde, wie zur flächendeckenden Evaluierung der Ästhetik von Landschaften neue textuelle Datenquellen benutzt werden können, die bisher nicht für die Landschaftsevaluierung verwendet wurden. Ausserdem wurde demonstriert, dass Text Mining einen neuen Blick auf die Wahrnehmung von Landschaften ermöglicht.

Abstract

The frequent use of terms such as urban sprawl and frugal land use is striking. The referendums on the Spatial Planning Act and Secondary-Home-Initiative recently carried out in Switzerland prove that natural landscapes are worth to be protected. At the same time they are something that is perceived as being threatened. But where are particularly sensitive landscapes and how does the spatial pattern of perceived landscape quality vary?

This work examines a new method for answering such questions. Most previous studies on the perception of landscape quality are based on surveys with limited spatial extent. However, new data on the Internet allow new approaches. In this work, two British data sets are used. One has the goal to document every square kilometer cell with a photograph and possibly a short text, the other contains landscape reviews that are based on a selection of images of the first data set. The aim of this work is to estimate the scenic value of landscapes based on texts by combining these two data sets. The motivation is that with such methods new textual data sources for the evaluation of landscapes can be exploited that have not been used before.

In a first step, the patterns in the landscape descriptions were revealed. It was found that terms of the natural areas are rated better than those of the urban space. This led to the awareness that to answer the second research question, namely the modeling of the landscape quality based on text analysis, a generic sentiment lexicon is inappropriate. Instead, specific landscape beauty lexicons were created and their performance was analyzed.

Several methods to estimate the landscape ratings were implemented. The best results were obtained with a single-word-treatment (unigrams) of the text. When the maximum voting variance was limited 53.2 % of the observed rating variance could be explained, the rank order of the ratings could be reproduced with an accuracy of 64.8 %. For unlimited voting variance 39.1 % of the rating variance was explained. The scenic value could be estimated more precisely when agreement on the rating was prevailing.

The contribution of this work is that it was shown how new textual data sources which have not been used for landscape evaluation can be used for the widespread estimation of the scenic value. Furthermore, it was demonstrated that text mining permits a new view on the perception of landscapes.

Inhalt

Abbildungen

Anmerkung: Die abgebildeten Karten sind, sofern nicht anders angegeben, genordet.

Tabellen

Glossar – Englische Begriffe

Beinn ist das schottische Wort für Berg.

Coire ist das gälische Wort für Kar, eine einem Amphitheater ähnliche Hohlform an einem Berghang, die durch Gletscher entstanden ist.

Corbett ist ein schottischer Ausdruck für Berge, die zwischen 2500 und 3000 Fuss hoch sind und eine relative Höhe von mindestens 500 Fuss aufweisen.

Crag ist ein felsiger Hügel, der durch Gletschererosion entstanden ist.

Glen bezeichnet ein Tal, das typischerweise lang und tief ist und oft durch Gletscher U-förmig erodiert wurde.

Loch ist das gälische Wort für See.

Lochan ist der Diminutiv für das gälische Wort *loch*.

Munro ist ein schottischer Ausdruck für Berge, die höher als 3000 Fuss sind.

Ravine bezeichnet ein Tobel.

Scree bezeichnet eine Schutthalde am Fuss von Felswänden (auch Talus genannt).

Tesco ist eine britische Supermarktkette.

Woolworth ist ein Handelsunternehmen.

1 Einleitung

1.1 Einführung ins Thema

Begriffe wie «Zersiedelung» und «haushälterische Bodennutzung», deren häufige Verwendung auffällt, und die in jüngster Zeit in der Schweiz durchgeführten Volksabstimmungen zum Raumplanungsgesetz und zur Zweitwohnungsinitiative zeugen davon, dass natürliche Landschaften als etwas Schützenswertes und gleichzeitig auch Bedrohtes wahrgenommen werden. Durch Siedlungsdruck, neue Energieerzeugungsformen, Ausbau von Verkehrswegen oder Tourismus entstehen vielfältige Ansprüche, die die Schönheit einer Landschaft bedrohen können. Doch welche Landschaften sind besonders wertvoll? Wo könnte beispielsweise der Bau eines Windparks oder eines Solarkraftswerks besonders negative Auswirkungen auf die Landschaftsqualität haben? Viele Forscher haben sich solchen Fragen angenommen, sich aber oft auf kleine Gebiete beschränkt.

Die vorliegende Arbeit verfolgt einen anderen Weg zur Behandlung dieser Frage. Durch das Wachstum des Internets sind viele nutzergenerierte Daten entstanden und öffentlich verfügbar geworden. Die Herangehensweise dieser Arbeit baut auf solchen nutzergenerierten Daten auf. Es wird zum einen auf einen Datensatz mit hunderttausenden von Landschaftsbewertungen zurückgegriffen, die durch ein Freiwilligenprojekt (ScenicOrNot) in Grossbritannien zustande gekommen sind. Zum anderen werden diese Landschaftsbewertungen mit Kommentaren verknüpft, die in einem anderen Datensatz (Geograph) zu den entsprechenden Landschaften vorliegen, um mit Methoden des Data Mining und der explorativen Datenanalyse zu untersuchen, welche Gesetzmässigkeiten und Muster in den Texten zu den Landschaften entdeckt werden können. In einem zweiten Schritt wird untersucht, wie mit maschinellem Lernen, Text Mining und computerlinguistischen Methoden eine Landschaftsbewertung aufgrund eines Textes geschätzt werden kann, ohne den Inhalt des entsprechenden Bilds zu berücksichtigen.

Die Motivation dieser Arbeit besteht darin, dass mit einer solchen Vorgehensweise neue, grosse Datenquellen – etwa Fotosammlungen im Internet – zur Evaluierung der Landschaftsästhetik erschlossen werden können, obwohl sie ursprünglich nicht dafür gedacht waren. Dadurch ermöglicht sich ein neuer Blick auf die Wahrnehmung von Landschaften. Ausserdem erlauben die in dieser Arbeit erforschten Methoden es auch, grossflächige Landschaftsevaluationen vorzunehmen ohne für die Bewertung Probanden rekrutieren zu müssen.

1.2 Forschungsfragen

Das Ziel dieser Arbeit ist zu erforschen, wie anhand von Landschaftsbeschreibungen die Schönheit einer textuell umschriebenen Landschaft geschätzt werden kann. Daraus ergeben sich die folgenden zwei Forschungsfragen:

Forschungsfrage 1:

Welche Muster kann man in den Kommentaren der Geograph-Bilder, die von ScenicOrNot bewertet wurden, entdecken?

Forschungsfrage 2:

Welcher Anteil der Varianz einer Landschaftsbewertung kann durch eine Analyse einer Bildlegende oder eines Begleittextes erklärt werden?

1.3 Einschränkung des Untersuchungs- und Themengebietes

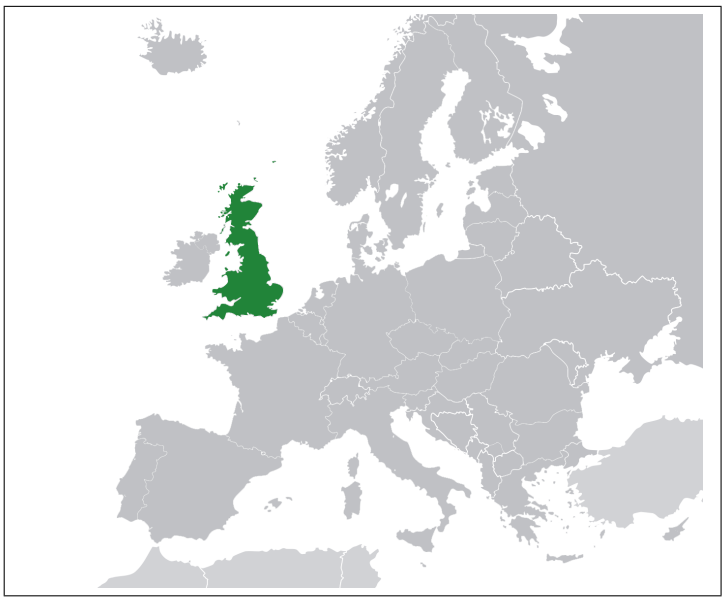


Abb. 1.1: Lage des Untersuchungsgebiets (grün eingefärbt) in Europa (Quelle: www.wikipedia.org)

Aufgrund der Datenverfügbarkeit richtet sich der Fokus dieser Arbeit auf das Gebiet Grossbritanniens und der Isle of Man und die verwendeten Methoden sind auf die englische Sprache zugeschnitten. Die Abb. 1.1 zeigt die Lage des Untersuchungsgebietes in Europa. Ferner baut die Arbeit zwar auf Landschaftsfotografien auf, der Inhalt derselben wird aber nicht analysiert und zur Analyse nicht benutzt. Die Bilder werden nur zu

Illustrationszwecken gezeigt. Das Ziel der Arbeit ist nicht, Fragen wie etwa die oben gestellten bezüglich einer optimalen Lage von Windpärken oder dergleichen zu beantworten, sondern die Untersuchung der Möglichkeiten, die die erwähnten Datensätze bieten.

1.4 Aufbau der Arbeit

Zu Beginn werden im Kapitel 2 die Hintergründe der für diese Arbeit relevanten Fachgebiete ausgeführt, welche folgende Bereiche umfassen: Landschaftsevaluierung, Opinion Mining und User generated content. Anschliessend werden die verwendeten Datensätze detailliert beschrieben (Kap. 3).

Im Kapitel 4 wird erörtert, wie sich die Erkenntnisse des wissenschaftlichen Hintergrundes auf die verwendeten Daten auswirken, und schliesslich werden bestehende Forschungslücken aufgezeigt. Das Vorgehen zur Beantwortung der Forschungsfragen wird im Kapitel 5 beschrieben und die damit erzielten Resultate werden im nächsten Kapitel (Kap. 6) behandelt. Im Kapitel 7 werden die Resultate diskutiert und die Arbeit schliesst mit einem Ausblick und den Schlussfolgerungen, die im Kapitel 8 zu finden sind.

2 Wissenschaftlicher Hintergrund

Die vorliegende Arbeit ist an der Schnittstelle diverser Fachgebiete, namentlich Landschaftsevaluierung (Abschnitt 2.2 auf der nächsten Seite), Opinion Mining (Abschnitt 2.3 auf Seite 11) und User generated content (Abschnitt 2.5 auf Seite 20) angesiedelt. In diesem Kapitel werden deshalb die Hintergründe dieser Gebiete erläutert.

2.1 Landschaftsbegriff und Landschaftsschönheit

«Landschaft» beziehungsweise das englische Wort «landscape», das mit dem deutschen in direkter Verbindung steht (*Jackson* 1986), wird in der Literatur gewöhnlich wie folgt definiert:

1. «An expanse of scenery that can be seen in a single view.»[1]

2. «A portion of the earth's surface that can be comprehended at a glance». (*Jackson* 1986)

Es geht also um einen Teil der Erdoberfläche, der in einem einzelnen Blick erfasst werden kann, was auch der alltäglichen Interpretation des Begriffs entspricht.

Landschaft setzt sich aus «Land» (*land*) und «Schaft» (*scape*) zusammen. Die urspüngliche Bedeutung von *Land*, welche auf die Goten zurückgeht, ist *gepflügtes Feld*. Bereits hier ist aber schon die auch heute gültige Bedeutung erkennbar, nämlich dass es sich um eine Fläche mit anerkannten Grenzen handelt. Die zweite Silbe, *Schaft*, bedeutete ursprünglich eine Sammlung ähnlicher Objekte – eine Landschaft ist demnach eine Sammlung von Ländern (im Sinn von gepflügten Feldern). (*Jackson* 1986)

Jackson (1986) bemerkt, dass der Landschaftsbegriff vor 1000 Jahren nichts mit *scenery* (in einem ästhetischen Sinn) zu tun hatte. Auch gemäss *Tuan* (1979) bezog sich der Begriff urspünglich auf ein Gut (*estate*). Erst später, ab dem 16. Jahrhundert, kam eine ästhetische Dimension hinzu, die sich nicht zuletzt dadurch manifestierte, dass das Kunstgenre der Landschaftsmalerei aufkam. Auch die englischen Landschaftsgärten, in denen die Natur ab dem 18. Jahrhundert idealisiert dargestellt wurde, können mit der gestiegenen Bedeutung der Ästhetik in Verbindung gebracht werden.

[1] Definition von TheFreeDictionary: http://www.thefreedictionary.com/landscape

Als Quintessenz kann festgehalten werden, dass die heutige ästhetische Konnotation dem Landschaftbegriffs nicht von Anfang innewohnte, sondern eine kulturelle Errungenschaft darstellt, die nicht zuletzt damit zu tun hat, dass der Mensch im Zug der industriellen Revolution gelernt hat, die Natur weitgehend zu kontrollieren (z. B. Flusskorrekturen, Sumpftrockenlegungen) (*Tuan* 1979). Dies hatte zur Folge, dass die Natur wesentlich weniger bedrohlich erschien.

2.2 Landschaftsevaluierung

Landschaftsevaluierung kann als das Vergleichen zweier oder mehr Landschaften in Bezug auf deren visuelle Qualitäten definiert werden (*Laurie* 1974). Die grundlegende Annahme bei der Evaluierung der visuellen Qualität einer Landschaft besteht darin, dass man davon ausgeht, dass Landschaften eine intrinsische, objektive Schönheit aufweisen (*Shuttleworth* 1979), die auch gemessen werden kann. Allerdings kommt Landschaftsschönheit durch zwei voneinander nicht zu trennende Quellen zustande: vom Objekt und vom Beobachter (*Arriaza et al.* 2004, *Laurie* 1974). Das bedeutet, dass die Schönheit erst im Geist konstruiert wird (*Tuan* 1979). Diese Sichtweise war in der Geschichte der Philosophie nicht unumstritten. Über Jahrhunderte wurde über zwei Paradigmen debattiert: Das objektive Paradigma, das besagt, die Schönheit eines Objekts sei im Objekt selbst inhärent; und das subjektive, das geltend macht, Schönheit entstehe erst im Betrachter. Mittlerweile besteht ein weitgehender Konsens, der das subjektive Paradigma bevorzugt (*Lothian* 1999).

Die systematische Erforschung der Schönheit von Landschaften hat in der Mitte der 1960er-Jahre begonnen (*Daniel et al.* 1976) und vor allem seit den 1970er-Jahren viel Aufmerksamkeit erfahren. Dabei werden verschiedene Methoden verwendet, die z. B. von *Arthur et al.* (1977) beschrieben wurden und nachfolgend erläutert werden.

2.2.1 Methoden zur Landschaftsevaluierung

Quantitative Methoden

Eine Möglichkeit, Landschaftsbewertungen zu erstellen, besteht darin, Laien Landschaften bewerten zu lassen. Hier wird das subjektive Paradigma der Schönheitsentstehung zugrunde gelegt und der Ansatz wird als wahrnehmungsbasiert beschrieben (*Daniel* 2001). Man geht davon aus, dass durch eine grosse Zahl bewertender Personen eine repräsentative Bewertung erzielt werden kann. Die Probanden bewerten die Landschaft auf einer numerischen Skala, die angibt, wie sehr sie die dargestellte Landschaft mögen (*Daniel et al.*

1976). Dabei werden die Probanden entweder vor Ort gebracht, was mit hohen Kosten und grossem zeitlichen Aufwand verbunden ist und trotzdem nur kleine Stichproben ermöglicht (*Roth* 2006), oder zur Bewertung werden Fotos der fraglichen Landschaften gezeigt, was wesentlich häufiger gemacht wird (*Schroeder* 1991). Hier stellt sich die Frage, ob Fotografien der Landschaften eine adäquate Repräsentation der Wirklichkeit sind, um die Landschaften evaluieren zu können. Verschiedene Autoren haben sich dieser Frage gewidmet, so zum Beispiel *Craik* (1972a) und *Stewart et al.* (1984). Die Resultate der Studien zeigten jeweils eine gute Übereinstimmung zwischen Bewertungen im Feld und Bewertungen von fotografierten Landschaften. *Stamps* (1990) hat eine Metaanalyse von elf Studien, die sich mit der Zulässigkeit von Fotos als Landschaftsrepräsentation befasst haben, durchgeführt. Sein Resultat ist eine Gesamtkorrelation von 0,86 zwischen den *in-situ*-Bewertungen und den Fotografien.

Es ist auch von Interesse, welche Landschaftselemente wie zur Gesamtbewertung beitragen. *Arriaza et al.* (2004) haben dies anhand von Fotos von Landschaften in Andalusien untersucht. Ihre Resultate zeigen, dass die wahrgenommene visuelle Qualität mit zunehmender Wildnis der Landschaft, der Präsenz von gut erhaltenen, menschgemachten Elementen, dem Prozentsatz der Pflanzenbedeckung, der Wassermenge im Bild, der Anwesenheit von Bergen und dem Farbkontrast steigt. Schon *Shafer und Mietz* (1970) haben ein Modell entwickelt, das die Landschaftsqualität aufgrund der in der Fotografie eingenommenen Fläche verschiedener Landschaftselemente schätzt. Durch die Analyse von Landschaftsmerkmalen haben *Schroeder und Daniel* (1981) ein statistisches Modell entwickelt, das aus Angaben wie Baumdichte, Bodenvegetation und Bauten die Landschaftsqualität schätzt.

Arthur et al. (1977) betonen, dass es beispielsweise für Landschaftsplaner nicht nur wichtig ist, wie eine Landschaft bewertet wird, sondern auch, wie die Bewertung zustande gekommen ist. Eine Möglichkeit dafür ist die oben beschriebene Analyse von *Arriaza et al.* (2004) der gezeigten Fotos, eine andere besteht darin, die Probanden verschiedene Landschaftselemente (z. B. *Dakin* (2003)) oder Adjektive (z. B. *Roth* (2006)) bewerten zu lassen, um eine differenzierte Beurteilung zu erhalten.

Gemäss *Schroeder* (1991) besteht ein Problem des quantitativen Ansatzes darin, dass verschiedene wichtige Fragen damit nicht beantwortet werden können, so zum Beispiel welche Erfahrungen die Menschen in bestimmten Landschaften machen, welche Werte mit bevorzugten Landschaften assoziiert werden oder wie wichtig diese Landschaften im Leben der Leute sind. Weiter ist aus quantitativen Analysen typischerweise nicht ersichtlich, wie Menschen auf eine Änderung einer spezifischen Landschaft reagieren. Um hier Antworten zu erhalten, ist ein offener, qualitativer Ansatz nötig.

Qualitative Methoden

Qualitative Methoden der Landschaftsevaluierung fokussieren auf die individuellen Erfahrungen einer Gegend, die Personen dort machen. Dazu werden offene Befragungsmethoden verwendet. *Schroeder* (1991), der quantitative und qualitative Methoden in einer Studie kombiniert hat, bat seine Teilnehmer, verschiedene Landschaften eines Arboretums verbal zu charakterisieren und anschliessend die Bedeutung der Landschaften zu beschreiben, indem sie ihre Gedanken, Gefühle und Erinnerungen daran erläuterten. Diese Aussagen wurden mit einer Inhaltsanalyse untersucht, um die Konzepte, die die Befragten in ihren mentalen Bildern haben, zu ergründen.

Auch die Arbeit von *Henwood und Pidgeon* (2001), die die symbolischen und persönlichen Assoziationen der Bewohner von Wales zu ihren Wäldern untersucht, verwendet sowohl quantitative als auch qualitative Elemente in ihrer Methodik. Im qualitativen Teil wurde in Gesprächen in Fokusgruppen erörtert, wie die Teilnehmer über die Wälder in ihrer Wohnregion denken, was den Bewohnern wichtig ist und welche Bedeutungen den Wäldern zugeschrieben werden. Die aufgezeichneten Gespräche wurden danach detailliert analysiert.

Eine weitere Möglichkeit zur qualitativen Datensammlung besteht darin, dass Teilnehmer während einer Reise Tagebücher führen, in welchen sie ihre Gedanken und Erlebnisse notieren. Dies wurde beispielsweise in der Studie von *Fredrickson und Anderson* (1999) gemacht, deren Ziel war, die Wildniserfahrung einer Reisegruppe zu erforschen. Die Tagebücher wurden mit einer Inhaltsanalyse untersucht und nach der Reise wurden zusätzlich mit allen Partizipanten Interviews geführt, wobei die Fragen offen gestellt wurden, um die Kandidaten zu ermutigen, ausführlich über ihre Erfahrungen zu berichten. In der Arbeit wurden die Interviews kurz nach der Reise durchgeführt, in einer ähnlichen Studie von *Heintzman* (2008) wurden die Befragungen erst 8 bis 10 Monate nach dem Trip durchgeführt. Die Interviews wurden jeweils transkribiert und ebenfalls mit einer Inhaltsanalyse untersucht.

Qualitative Studien haben typischerweise wenig Teilnehmer – bei *Heintzman* (2008) beispielsweise nur 6, bei *Schroeder* (1991) immerhin 29 – und decken oft nur relativ kleine Regionen ab. Dafür ermöglichen sie ein vertieftes Verständnis für die Meinungen über Landschaften.

An der Schnittstelle zwischen qualitativen und quantitativen Methoden können deskriptive Inventuren gesehen werden.

Deskriptive Inventuren In beschreibenden Inventuren werden Landschaftselemente erst identifiziert, dann beschrieben und allenfalls bewertet. Deskriptive Inventuren können sowohl qualitative als auch quantitative Resultate ergeben.

Um quantitative Bewertungen zu erhalten, werden die einzelnen Landschaftskomponenten quantifiziert, so dass sie gewichtet, verglichen und aggregiert werden können (*Arthur et al.* 1977).

Deskriptive Inventuren beschreiben, welche Landschaftselemente zur Schönheit einer Landschaft beitragen. Gemäss *Arthur et al.* (1977) erhöht beispielsweise landschaftliche Variation die Bewertung. *Cherem* (1973) hat in einer damals innovativen Studie Wanderer mit Kameras ausgerüstet und danach die fotografierten Landschaften analysiert. Dabei stellte er fest, dass die Wanderer «komplexe» Landschaften mit viel Abwechslung bevorzugten. Öfter werden Inventuren aber von Experten erstellt, deren Bewertung nicht notwendigerweise mit jener der Öffentlichkeit übereinstimmen muss. Gemäss *Daniel* (2001) wurde der expertenbasierte Ansatz vor allem in der Praxis des Landschaftsmanagements angewendet. Es wird kritisiert, dass die Landschaftsqualität oft nur in wenigen Klassen beschrieben wird (bspw. *tief*, *mittel*, *hoch*) und dass zwischen verschiedenen Experten oft unterschiedliche Bewertungen resultieren. Dem expertenbasierten Ansatz liegt das oben beschriebene objektive Paradigma zugrunde (*Daniel* 2001).

2.2.2 Internetbasierte Landschaftsevaluierung

Bei der traditionellen Art und Weise, Landschaften mit Fotografien zu bewerten, werden Probanden (oftmals Studenten) in ein Labor eingeladen, um dort die Fragebögen zu beantworten. Für die Studie von *Hunziker und Kienast* (1999), die die Auswirkung von Vergandung auf die Landschaftsqualität untersucht haben, wurden beispielsweise 181 Studenten befragt. Die Autoren merken selbst an, dass die empirische Basis vergrössert werden müsste und dass Personen aus ganz Europa vertreten sein müssten, um wirklich repräsentative Resultate zu erzielen. Genau dies ist aber mit dieser Methode schwierig (*Roth* 2006).

Das Internet bietet die Möglichkeit, ein grösseres Publikum zu erreichen, das im Prinzip über ein beliebig grosses Territorium verteilt sein kann. *Wherrett* (1999) hat als einer der ersten die Thematik der internetbasierten Landschaftsevaluierung erörtert. Einige der damaligen Probleme (limitierte Farbdarstellung, Geschwindigkeit der Verbindung, Bildschirmgrösse) sind mittlerweile kaum noch relevant. *Wherrett* (1999) spricht auch Probleme über das erreichbare Publikum an. Zwar hat sich der Anteil der Menschen mit Internetzugang seit damals stark vergrössert, aber trotzdem sind die Antworten nicht zwingend repräsentativ für die gesamte Bevölkerung, sondern eher für die internetnutzende Bevölkerung (*Wherrett* 1999). Es gibt keine Kontrolle darüber, wer die Fragebögen ausfüllt, aber durch das Internet ist es relativ einfach, das soziodemografische Profil festzustellen, wenn man einen strukturierten Fragebogen vorlegt.

2.2.3 Beprobung

Bei der Datensammlung für eine Landschaftsevaluierung stellt sich die Frage, von welchen Punkten aus die Fotos gemacht werden sollen, die später zur Bewertung verwendet werden. Verschiedene Forscher (z. B. *Anderson und Schroeder* 1983 und *Buhyoff et al.* 1986) haben als Samplingmethode ein zufälliges Verfahren verwendet, um die Standpunkte in der Region zu bestimmen. Gemäss *Hull et al.* (1989) besteht eine weitere Beprobungsstrategie darin, Punkte auszuwählen, von denen aus die betreffende Landschaft oft betrachtet wird und die deshalb repräsentativ sind. Andere Autoren (z. B. *Schroeder und Daniel* 1981) wählen ihre Samplingpunkte so, dass die sichtbaren Landschaftselemente repräsentativ für die zu erforschende Region sind. Die vierte von *Hull et al.* (1989) beschriebene Strategie sieht vor, dass die Fotopunkte so selektiert werden, dass eine Forschungshypothese getestet werden kann. Das bedeutet, dass beispielsweise eine gewisse Baumdichte oder ein bestimmter Kontrast auf dem Foto zu sehen sein soll. Diese Studien verfolgen in der Regel nicht das Ziel, eine repräsentative Bewertung einer Region zu erstellen. Wenn der Fotopunkt bestimmt ist, stellt sich die Frage, in welche Richtung das Foto geschossen werden soll.

Hull et al. (1989) beschreibt für diese Problematik verschiedene Methoden, die in anderen Studien angewendet wurden. *Daniel et al.* (1976) etwa haben die Richtung des ersten Fotos zufällig bestimmt und die drei nachfolgenden in Winkelabständen von jeweils 90° geschossen. Die Landschaftsqualität entspricht dann dem Durchschnitt der vier bewerteten Szenen. *Schroeder und Daniel* (1981) haben in einer Studie über die Landschaftsqualität, die von Strassen aus gesehen werden kann, mit vorgegebenen Winkeln zur Strasse gearbeitet. Laut *Hull et al.* (1989) haben viele Studien rigorose Einschränkungen, die zur Folge haben, dass auf den abgelichteten Landschaftsausschnitten lediglich ein Baumstamm oder ein Gebäude zu sehen ist, was kaum der menschlichen Wahrnehmung einer Landschaft entspricht. *Buhyoff et al.* (1986) argumentiert deshalb, dass bei der Auswahl des fotografierten Landschaftsausschnittes Flexibilität nötig ist und erlaubt beim Fotografieren eine rechtwinklige Verschiebung mit festgelegtem Aufnahmewinkel vom Ausgangspunkt aus, um eine freie Sicht auf die Landschaft zu ermöglichen.

Schliesslich können Fotos von zu evaluierenden Landschaften auch von Laien geschossen werden, was von *Hull et al.* (1989) als «participant photography» bezeichnet wird. Der Vorteil besteht darin, dass so häufig besuchte Punkte vermehrt fotografiert werden, was der Repräsentativität zuträglich ist, und gleichzeitig Szenen fotografiert werden, die für die Besucher von Bedeutung sind (*Hull et al.* 1989).

2.2.4 Zeitliche Veränderung der Landschaft

Die Akzeptanz für die Verwendung von Fotos zur Landschaftsbewertung (siehe S. 6) impliziert, dass Landschaft etwas Statisches darstellt, doch das Gegenteil ist der Fall. *Hull und McCarthy* (1988) haben die Veränderungen, die sich in einer Landschaft abspielen können kategorisiert. Es gibt langsame, schwierig zu erkennende Wechsel, wie zum Beispiel das Vegetationswachstum und plötzliche Veränderungen, wie ein Wetterumschwung. Es gibt regelmässige Änderungen wie die saisonale Veränderung der Vegetation oder Sonnenuntergänge und häufige Wechsel wie das Aufkommen von Wind und seltene Ereignisse, beispielsweise Überflutungen. Schlussendlich kann man die Länge diverser Vorgänge unterscheiden, von Langzeitereignissen über mittelfristige Veränderungen bis hin zu kurzzeitigen Phänomenen wie etwa dem Auftauchen von wild lebenden Tieren. *Hull und McCarthy* (1988) haben den Einfluss von wild lebenden Tieren auf Landschaftsfotografien auf die Bewertung der Landschaft untersucht. Dabei zeigte sich, dass die Landschaften positiver bewertet wurden, wenn Tiere darauf zu sehen waren. Sogar schon die blosse Erwartung, dass Tiere gesehen werden könnten, erhöht die Landschaftsbewertung.

Malm et al. (1981) haben die Wahrnehmung der Luftqualität analysiert. Insbesondere bei wolkenlosem Himmel verschlechtern Aerosole in der Luft die Landschaftsbewertung, während dieser Effekt bei bedecktem Himmel geringer ist. Ein weiterer regelmässiger Wechsel in vielen Gegenden sind die saisonalen Veränderungen der Vegetation. *Buhyoff und Wellman* (1979) haben erforscht, wie die Bewertung einer saisonal veränderlichen Landschaft mit dem Zeitpunkt der Bewertung zusammenhängt, und herausgefunden, dass Herbstbilder im Herbst besser bewertet werden als dieselben Bilder in anderen Jahreszeiten.

Auch die Geräuschkulisse hat bei der Landschaftsevaluierung eine Bedeutung. Die Auswirkungen von anthropogenem Lärm (Flugzeug- und Verkehrslärm sowie Stimmen) und Naturgeräuschen (Vogelgezwitscher und Wind im Blattwerk) wurden von *Benfield et al.* (2010) analysiert. Die natürlichen Geräusche hatten wenig bis keinen Effekt, aber die menschlichen Geräusche haben die wahrgenommene Landschaftsqualität verschlechtert. *Herzog* (1985) hat untersucht, welche Formen von sich bewegendem Wasser – ebenfalls ein sich veränderndes Landschaftselement – in der Landschaft bevorzugt werden. Seine Ergebnisse zeigen, dass schnell fliessende Gewässer in gebirgigen Regionen am stärksten bevorzugt werden, während sumpfige Gegenden am schlechtesten abschneiden.

2.2.5 Einfluss von Landschaftsbeschreibungen auf Bewertungen

Hodgson und Thayer (1980) haben ein Experiment durchgeführt, in dem 15 Bilder von Landschaften nach Schönheit geordnet werden sollten. Bei einer Versuchsgruppe

waren 4 der 15 Bilder mit naturbezogenen Begriffen (lake, pond, stream bank, forest growth) beschriftet, bei einer anderen Gruppe waren dieselben Bilder mit Begriffen, die menschlichen Einfluss implizieren, versehen (reservoir, irrigation, road cut, tree farm). Abgesehen von den Beschriftungen waren die Bilderreihen identisch. Es zeigte sich in allen Experimenten, dass Bilder mit naturbezogenen Begriffen höhere Ränge erzielten als solche, die menschlichen Einfluss suggerierten. Die Landschaftsbewertung hängt also nicht nur mit der Landschaft selbst zusammen, sondern auch mit der Erwartungshaltung und Meinung des Bewertenden. Das stützt wiederum das subjektive Paradigma und die Vorstellung, dass Schönheit erst im Geist konstruiert wird.

2.3 Opinion Mining

Das Ziel von *Opinion Mining* ist, aus unstrukturiertem, natürlichen Text mithilfe von Computerprogrammen automatisch die Stimmung (positiv, negativ oder neutral) zu extrahieren (*Pang und Lee* 2008). Ein wesentliches Merkmal von Opinion Mining ist, dass es nicht darum geht zu erkennen, wovon ein Text handelt, sondern darum, die Stimmung unabhängig vom behandelten Thema zu erkennen (*Esuli und Sebastiani* 2006). Allerdings fällt auf, dass viele Ansätze sich jeweils auf wenige Themengebiete konzentrieren, weil oft eine themenspezifische Sprache verwendet wird und Mehrdeutigkeiten so besser vermieden werden können.

2.3.1 Geschichte

Gemäss *Pang und Lee* (2008) können die Arbeiten von *Carbonell* (1979) und *Wilks und Bien* (1983) als frühe Vorboten der Stimmungserkennung bezeichnet werden. Die eigentlichen Anfänge des Opinion Mining reichen aber in die Mitte der 1990er-Jahre zurück. *Sack* (1995) präsentiert ein System, das mit Hilfe einer *actor-role*-Analyse den in einer Zeitung beschriebenen Standpunkt eines Akteurs (*actor*) aufgrund seiner Rolle (*role*) herausfinden soll. Die einzelnen Akteure und Rollen sowie die Standpunkte müssen von Hand in einer Datenbank eingegeben werden. Eine Software analysiert dann einen Zeitungsartikel und bestimmt den wahrscheinlichsten Standpunkt.

Terveen et al. (1997) haben eine Suchmaschine (Phoaks) entwickelt, die Usenet-Beiträge nach Empfehlungen weiterer Websites durchsucht. Dazu werden in einem ersten Schritt Beiträge mit URLs gesucht. Die gefundenen URLs müssen dann kategorisiert werden – beispielsweise sollten Empfehlungen interessanter Websites von persönlichen Homepages getrennt werden. Dazu verwenden die Autoren syntaktische Regeln. Die Autoren berichten von hunderten Bedingungen, die analysiert werden, um die gefundenen URLs schliesslich

16 Kategorien zuzuteilen. Die genauen Regeln werden aber nicht besprochen. Obwohl es im Paper nicht explizit erläutert wird, ist ersichtlich, dass eines der Ziele dieses Projekts darin bestand, subjektive Beiträge (persönliche Homepages) von objektiven Inhalten (URL-Weiterempfehlungen) zu trennen.

Auch *Wiebe* (2000) verfolgt das Ziel der Trennung von subjektiven und objektiven Inhalten. Subjektivität in natürlicher Sprache bezieht sich auf Aspekte der Sprache, die Meinungen und Bewertungen ausdrücken können. Neben automatischen Zusammenfassungen, die gemäss *Wiebe* (2000) möglichst nur Fakteninformation beinhalten sollen, erwähnt er als mögliche Anwendungsgebiete auch das Suchen von Produkte-Reviews in Onlineforen. Dieses Gebiet haben später viele Autoren erforscht.

Subjektivität

Ein wichtiger Hinweis, ob ein Satz subjektiv ist, ist die Anwesenheit von Adjektiven. Wenn ein Satz mindestens ein Adjektiv enthält, ist die Wahrscheinlichkeit, dass er subjektiv ist, 55.8% (*Bruce und Wiebe* 1999). Darauf basierend hat *Wiebe* (2000) einen probabilistischen Klassifikator erstellt, der die Subjektivität eines Satzes mit einer Wahrscheinlichkeit von 72% erkennt. Er hat dazu Zusatzinformation über Adjektive verwendet, namentlich die Polarität der Adjektive. Die Polarität eines Adjektivs bezeichnet eine negative bzw. positive Semantik. *Wiebe* (2000) hat auf der Arbeit von *Hatzivassiloglou und McKeown* (1997) aufgebaut. Diese beiden Autoren haben ausgehend von einigen manuell klassierten Adjektiven die Polarität vieler Adjektive bestimmt. Sie haben dazu die Konjunktionen (z.B. *und*, *aber*) zwischen Adjektiven in einem Text-Korpus mit 21 Millionen Wörtern des *Wall Street Journal* analysiert. Konjunktionen sind ein guter Indikator für die semantische Orientierung der Adjektive, die sie verbinden: *und* verbindet normalerweise Adjektive gleicher Orientierung (*z. B. gerecht und gut*). *Gerecht und schlecht* wäre eine ungewöhnliche Verwendung der Adjektive. Für *aber* gilt das Gegenteil: Es verbindet üblicherweise Adjektive unterschiedlicher Orientierung. Die Konjunktionen geben aber keinen Aufschluss über die eigentliche semantische Orientierung der verbundenen Adjektive. Die Adjektive werden durch einen Clustering-Algorithmus in zwei Gruppen eingeteilt, die schliesslich als positiv und negativ gekennzeichnet werden.

Mit der Kombination von Subjektivitätserkennung und Polaritätsbestimmung wurde dann versucht, aus Produkte-Reviews von Laien, die im Internet zahlreich zu finden sind, die vorherrschenden Meinungen zu bestimmten Produkten zu finden. So haben *Morinaga et al.* (2002) ein Framework implementiert, das zu bestimmten Produkten die Meinungen aus dem Internet extrahiert und die Wahrscheinlichkeit angibt, dass es sich beim Statement auch um eine Meinung handelt.

Opinion Mining einzelner Aspekte

Wie *Yi et al.* (2003) beschreiben, wurde bei den frühen Ansätzen des Opinion Mining über Produkte oft versucht, ein gesamtes Dokument (Review) als positiv, negativ oder irgendwo dazwischen zu klassieren. *Yi et al.* (2003) präsentieren eine Methode, die die Meinungen zu einzelnen Aspekten eines Produktes erkennen soll. Die beschriebenen Aspekte eines Produkts sollen automatisch erkannt werden. Dazu wird die Tatsache verwendet, dass die beschriebenen Aspekte jeweils Nomen sind, die zusätzlich üblicherweise mit *the* (bestimmter Artikel) beginnen. Durch die Anwendung statistischer Methoden konnten die Autoren dadurch eine sehr gute Trefferquote erzielen. Danach wird die Meinung zu den einzelnen Produktefeatures bestimmt. Dazu wird ein *sentiment lexicon* verwendet, in dem die Polarität von ca. 2500 Adjektiven und fast 500 Nomen verzeichnet ist. Das Ergebnis der beschriebenen Methode ist die Einstellung der Konsumenten zu einzelnen Produktefeatures, von denen in einem Satz auch mehrere vorkommen können, anstelle eines einzigen Ratings für ganze Sätze oder ganze Texte.

Auch *Nasukawa und Yi* (2003) haben eine Methode entwickelt, die nicht die Stimmung eines gesamten Dokuments erfassen soll, sondern die Meinung zu einem bestimmten, vorgegeben Subjekt (z. B. *Range Rover*). Dazu wird eine Syntaxanalyse (*syntactic parser*), also die Erkennung von Subjekt und Objekt, mit einem *syntactic lexicon* kombiniert. Im Lexikon ist neben den Wörtern und deren Polarität auch vermerkt, ob sich das entsprechende Wort und dessen Polarität auf ein Subjekt oder ein Objekt bezieht. Im Lexikon sind Adjektive, Adverbien, Nomen und Verben enthalten. Die Autoren haben sich auf die Erkennung von semantischen Beziehungen konzentriert und berichten, dass sie je nach verwendeter Datengrundlage eine Präzision zwischen 75 und 95% erreicht haben.

Bewertende Wortgruppen

Whitelaw et al. (2005) widmen sich einem anderen Aspekt der Meinungsäusserung, nämlich den sogenannten *appraisal groups*. Hier geht es darum, dass Meinungen oft nicht in einzelnen Wörtern geäussert werden, sondern in bewertenden Wortgruppen (*appraisal groups*), beispielsweise «not really very good».

Es wurde eine Taxonomie der *appraisal groups* erstellt, welche in Abbildung 2.1 auf der nächsten Seite dargestellt ist. Die Taxonomie hat vier Aspekte:

- *Attitude*

  Attitude beschreibt die Art der Bewertung: Entweder persönlich (z. B. «happy»), anerkennend (z. B. «ugly», die Art muss nicht notwendigerweise positiv sein) oder eine soziale Beurteilung (z. B. «idiotic»).

- *Orientation*

  Die Orientierung zeigt an, ob eine Bewertung positiv oder negativ ist.

- *Graduation*

  Mit graduation wird die Intensität der Bewertung bezeichnet. So verstärkt z. B. «very» ein Statement, während es durch «slightly» abgeschwächt wird.

- *Polarity*

  Polarity bezeichnet bei *Whitelaw et al.* (2005) die Anwesenheit einer Negation (z. B. «not»), welche die Orientierung umkehrt.

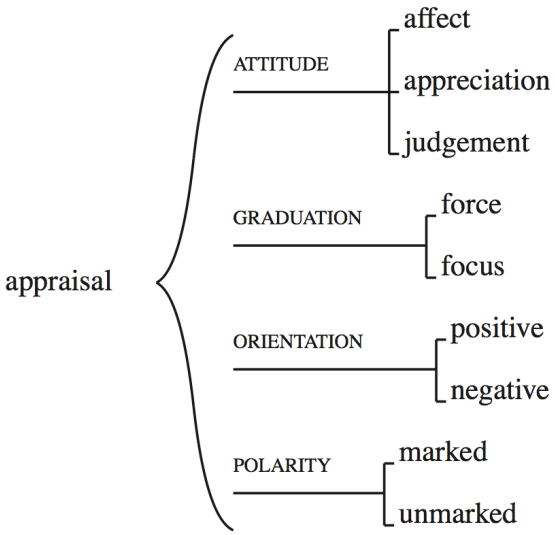


Abb. 2.1: Hauptattribute von *appraisal* (aus *Whitelaw et al.* 2005)

Gemäss den Autoren fokussierten die meisten Ansätze des Opinion Mining bis dahin auf die *Orientation*. Mit diesen Überlegungen erstellten sie ein Lexikon dieser *appraisal groups* und testeten es mit jeweils 1000 negativen und 1000 positiven Filmbeurteilungen der IMDb[2]. Die Präzision war laut den Autoren sehr gut (90.2%). Die wichtigsten Adjektive in *appraisal groups* kamen aus der Gruppe der *appreciation*, also anerkennende Wortgruppen (sowohl positiv als auch negativ), was bei Filmbewertungen auch nicht allzu überraschend ist.

Auch *Polanyi und Zaenen* (2006) haben sich mit der Bedeutung von modifizierenden Wörtern auseinandergesetzt und argumentieren, dass zur Beurteilung, ob ein Statement positiv oder negativ ist, nicht nur jedes Wort zu betrachten und zu bewerten ist, sondern

[2]Internet Movie Database (www.imdb.com)

sein Kontext berücksichtigt werden muss. *Taboada et al.* (2011), die die Stimmungsanalyse im Gegensatz zu *Whitelaw et al.* (2005) anwendungsorientiert angehen, nennen diese Problematik *intensification* und präsentieren eine Berechnungsmethode zur Bestimmung der Werte für modifizierte Wörter (z. B. *really fantastic*), die auf einer prozentualen Verstärkung des ursprünglichen Wertes beruht. Damit werden emotionsstarke Wörter mehr verstärkt als schwache Wörter.

Für die Negation schlagen *Taboada et al.* (2011) einen sogenannten *polarity shift* vor. Dabei wird ein fixer Betrag subtrahiert (positives Wort negiert) bzw. addiert (negatives Wort negiert), wobei die Stimmungswerte auf einer symmetrischen Skala (z. B. –5 - +5) vorliegen. Die Alternative, die *Taboada et al.* (2011) beschreiben, ist der *negation switch*, bei dem der ursprüngliche Wert mit –1 multipliziert wird. Diese Herangehensweise wird von den Autoren aber als grundsätzlich fehlerhaft beschrieben.

2.3.2 Social Media

MySpace

Mit dem Aufkommen von Social Media wurde die Erforschung von Opinion Mining auch auf diese Gebiete ausgedehnt. *Thelwall et al.* (2010) haben sich der Emotionsdetektion im sozialen Netzwerk MySpace[3] gewidmet. Die Herausforderung in diesem Gebiet von informellem Text besteht darin, dass oft Rechtschreib- und Grammatikregeln ignoriert werden und viel Slang verwendet wird, was zur Folge hat, dass Standardalgorithmen Mühe mit diesen Daten haben. *Thelwall et al.* (2010) haben einen Algorithmus erstellt, der auf diese Eigenheiten von sozialen Netzwerken eingeht. Einerseits ist eine Rechtschreibkorrektur eingebaut, die die Wiederholung von einzelnen Buchstaben korrigiert (so wird z.B. aus «helllo» «hello»), allerdings werden nicht alle wiederholten Buchstaben korrigiert, denn gemäss den Autoren ist die Wiederholung von Buchstaben eine verbreitete Art, um Emotionen zu verstärken (z.B. «hooooooooot»). Dies wird vom Algorithmus berücksichtigt, und entsprechende Wörter werden stärker positiv oder negativ gewichtet. Daneben wird eine Liste mit «Booster»-Wörtern verwendet, die Wörter abschwächt («some») oder verstärkt («very» etc.) und bei Negationen wird die Polarität umgekehrt. Weiter wird auch die Verwendung von Ausrufezeichen und Emoticons (z.B. «:-)») berücksichtigt und negative Emotionen werden in Fragen ignoriert. Der entwickelte Algorithmus, SentiStrength, wurde später weiter verbessert und in Version 2.0 vorgestellt (*Thelwall et al.* 2012).

[3] www.myspace.com

Twitter

Pak und Paroubek (2010) haben Twitter[4] als Grundlage für eine linguistische Analyse und das Training eines Emotionsklassifikators verwendet. Es wurden positive und negative Tweets gesammelt, indem nach Tweets mit positiven und negativen Emoticons gesucht wurde. Zusätzlich wurden neutrale Tweets gesammelt, indem 44 Konten von Zeitungsredaktionen abgefragt wurden. Weil ein Tweet nicht länger als 140 Zeichen lang sein kann, gehen die Autoren davon aus, dass jeweils nur ein Satz vorkommt, der eine einzige Emotion überbringt. Die Resultate der linguistischen Analyse sind in der Abbildung 2.2 und der Abbildung 2.3 auf der nächsten Seite dargestellt.

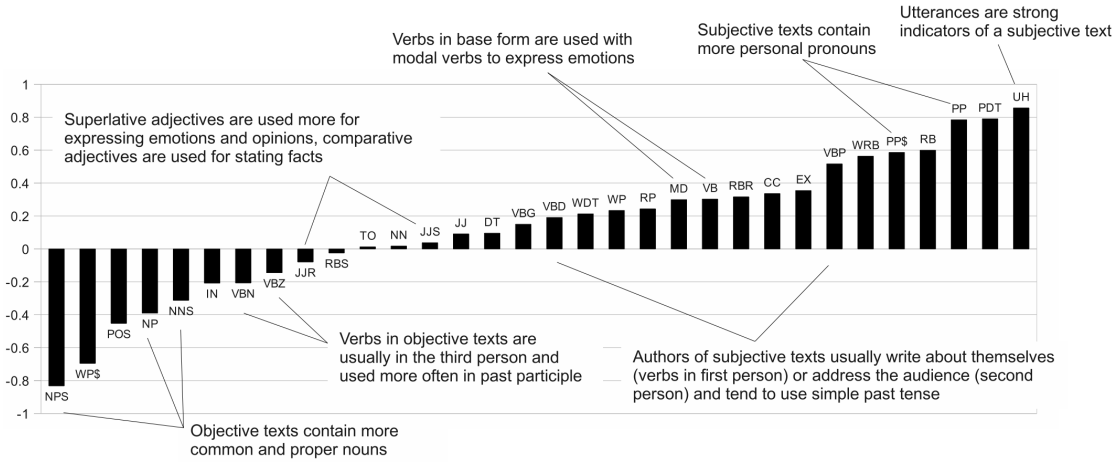


Abb. 2.2: Wortarten in objektiven und subjektiven Texten (aus *Pak und Paroubek* 2010)

Abbildung 2.2 zeigt, welche Wortarten eher in objektiven (links) und subjektiven (rechts) Texten vorkommen. So sind Ausrufezeichen und persönliche Pronomen (z. B. ich, mich, mein) starke Indikatoren für subjektiven Text und Nomen kommen häufiger in objektiven Texten vor (dies stimmt mit den Beobachtungen von *Esuli und Sebastiani* (2006) überein). Die Bestimmung der Wortart kann also Hinweise auf die Art des Textes geben.

Abbildung 2.3 auf der nächsten Seite zeigt, welche Wortarten vermehrt in positiven (links) und negativen (rechts) Tweets verwendet werden. So enthalten negative Texte zum Beispiel öfter Verben in einer Vergangenheitsform.

Flickr – Opinion Mining mit Bildern

Kisilevich et al. (2010) haben Kommentare, die zu Fotos auf der Plattform Flickr[5] abgegeben wurden, analysiert. Die Kommentare können sich einerseits auf das Foto und

[4] www.twitter.com
[5] www.flickr.com

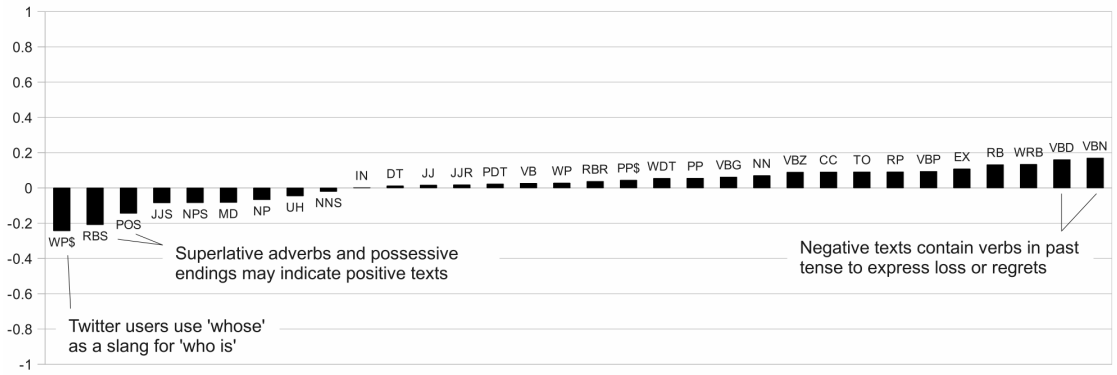


Abb. 2.3: Wortarten in positiven und negativen Texten (aus *Pak und Paroubek* 2010)

dessen Qualität beziehen («great shot»), andererseits aber auch auf das abgebildete Motiv. Die Autoren verfolgen deshalb einen Ansatz mit dem Ziel, diese beiden Aspekte zu trennen und separat von einander zu bewerten. Um die Trennung vorzunehmen, werden Wörter definiert, die sich auf sogenannte «photo features» (etwa *great shot*) beziehen – alle anderen Sätze beziehen sich dann auf Äusserungen, die das Motiv betreffen. In dieser Studie werden die Kommentare nicht nur binär zwischen «positiv» und «negativ» unterschieden, sondern jedem Foto wird eine Zahl auf einer kontinuierlichen Skala für die Meinungen über das Foto selbst und das abgebildete Motiv zugewiesen. Die Orientierung der Wörter wird einem manuell erweiterten Lexikon (Internet General Inquirer Lexicon) entnommen und die Stärke der gefühlsausdrückenden Adjektive ist umso grösser, wenn sie seltener in den Kommentaren verwendet werden. Das bedeutet, dass häufig verwendete Wörter in der Lesart der Autoren eine kleinere Emotion ausdrücken.

Siersdorfer et al. (2010) haben *sentiment analysis* ebenfalls auf Flickr angewendet. Um Bilder mit positiven und negativen Emotionen zu sammeln, sind die Autoren wie folgt vorgegangen: Es wurden die 1000 positivsten und die 1000 negativsten Wörter des unten beschriebenen SentiWordNet verwendet, um nach Bildern in Flickr zu suchen. So wurden insgesamt über eine halbe Million Bilder heruntergeladen, deren visuelle Merkmale, insbesondere das globale und das lokale Farbhistogramm, analysiert wurden. Mit diesen Informationen können die Emotionen und die Histogramme korreliert werden. Sowohl das globale wie auch das lokale Farbhistogramm zeigt dabei denselben Trend: Fotos mit positiven Wörtern werden von erdfarbenen Tönen und Hautfarben dominiert, negative Wörter weisen vermehrt Grün- und Blautöne auf. Dieses Resultat spiegelt die Wahrnehmung von warmen (positiv) und kühlen (negativ) Farben.

2.3.3 Zusammenfassung Opinion Mining

Um die Stimmung eines Textes erkennen zu können, muss zuerst bestimmt werden, ob es sich überhaupt um einen subjektiven Text handelt, oder ob Fakten beschrieben sind. Falls ein subjektiver Text vorliegt, muss entschieden werden, ob der Text als ganzes klassiert werden soll, eine satzweise Erkennung stattfinden soll oder ob sogar die einzelnen Subjekte erkannt werden sollen und für jedes der erkannten Subjekte eine Meinung berechnet werden soll. Bei der erstgenannten Vorgehensweise wird typischerweise die vorherrschende Meinung bestimmt, die anderen Methoden erlauben die Koexistenz negativer und positiver Emotionen und schliesslich ist es möglich, die Emotionen nicht nur binär, sondern graduell (je nach Stärke) zu erkennen.

Für die eigentliche Bestimmung der Meinung bietet sich die Verwendung eines *sentiment lexicon* an, das Wörtern eine negative, positive oder neutrale (objektive) Bewertung zuweist. Weiter können die Wörter der Sätze nicht einzeln angeschaut werden, sondern sie müssen in ihren Kontext gestellt werden, um beispielsweise eine Negation zu erkennen oder die Verwendung von Modifikatoren («very») zu berücksichtigen. Weitere Probleme ergeben sich durch mehrdeutige Wörter, deren Bedeutung sich erst im Kontext erschliesst. Schliesslich ist zu beachten, für welches Themengebiet die Stimmungserkennung bestimmt ist und diese entsprechend den Eigenheiten dieses Gebietes anzupassen. So können etwa Social Media-Plattformen nicht auf dieselbe Weise wie Produkte-Reviews analysiert werden. Diese Domain-Spezifität wird auch von *Taboada et al.* (2011) beschrieben.

2.3.4 Umfassendes *sentiment lexicon*

Verschiedene bereits erwähnte Autoren haben Wortlisten für positive und negative Wörter verwendet, die aber typischerweise manuell erstellt wurden. Solche Stimmungslexika sind für den lexikonbasierten Ansatz der Stimmunserkennung nötig. *Esuli und Sebastiani* (2006) haben auf der Basis von WordNet 2.0[6] eine umfassende Datenbank erstellt, die für jedes Synset einen Wert zwischen 0 und 1 für Negativität, Positivität und Objektivität vergibt, wobei die drei Werte in der Summe 1 ergeben. Jedes Synset kann für jeden der drei Werte eine Zahl ungleich 0 haben. Das bedeutet, dass solche Wörter sowohl in positivem wie in negativem Kontext verwendet werden können. Ausgehend von einem kleinen, manuell erstellten Seed-Set wurden durch automatische Klassifikatoren allen Synsets Werte zugewiesen. Dabei enstand SentiWordNet 1.0. In einem weiteren Paper (*Baccianella et al.* 2010) wird SentiWordNet 3.0 vorgestellt, das WordNet 3.0 als Grundlage hat und sich durch eine verfeinerte Berechnung der Werte auszeichnet.

[6] WordNet ist eine Datenbank, die semantisch verwandte, englische Wörter (Synonyme) in Gruppen fasst, die Synset genannt werden und für jedes Wort eine kurze Definition beinhaltet.

Eine Analyse der Klassifikation zeigt, dass nur wenige Terme völlig eindeutig sehr negativ oder sehr positiv sind. 75.37% der Terme haben eine Objektivität von 1, während nur 0.01% der Wörter eine Positivität oder Negativität von 1 erreichen. Weiter zeigt sich, dass vor allem die Wortarten «Adjektive» und «Adverbien» wenigstens teilweise subjektiv sind (39.66% bzw. 35.7%), Nomen und Verben hingegen sind nur in 9.98% bzw. 11.04% der Fälle subjektiv. Offenbar wird Meinung in natürlicher Sprache vor allem durch modifizierende Wörter (Adjektive, Adverbien) vermittelt. SentiWordNet ist für die wissenschaftliche Forschung frei zugänglich und hat Werte für 117'659 Synsets.

Beide verfügbaren Versionen von SentiWordNet wurden für Vergleichszwecke von *Taboada et al.* (2011) mit verschiedenen Methoden zur Transformierung der Stimmungswerte evaluiert. Die besten Werte wurden mit der Version 3.0 erzielt, wobei für Wörter mit verschiedenen Bedeutungen ein Durchschnitt berechnet wurde.

2.4 Wie wird Sprache verwendet?

Zipf (1935) argumentiert in seinen linguistischen Arbeiten, dass komplexere Sprachelemente seltener vorkommen. So nehme die durchschnittliche Länge eines Wortes ab, je häufiger das betreffende Wort in der Sprache verwendet wird. Dies ist ein Beispiel für eines seiner Haupterkenntnisse: Das Prinzip des geringsten Aufwandes (*principle of least effort*). *Piantadosi et al.* (2011) hingegen argumentieren, dass sich zur Vorhersage der Verwendungshäufigkeit eines Wortes der Informationsgehalt eines Wortes besser eignet als dessen Länge.

Neben dem Zusammenhang zwischen Verwendungshäufigkeit und Wortlänge ist vor allen Dingen auch von Interesse, wie Emotionen in Sprache übermittelt werden. Gemäss der *Pollyanna-Hypothese* (*Boucher und Osgood* 1969) werden emotional positive Wörter häufiger verwendet als negative Wörter. Einfach gesagt bedeutet dies: *humans tend to look on (and talk about) the bright side of life* (*Boucher und Osgood* 1969). Auch *Garcia et al.* (2011) unterstützen die Hypothese, dass es bei den menschlichen Äusserungen eine positive Verzerrung gibt. Als Datenquelle wurde das *n-gram dataset* von Google[7] verwendet, das gemäss den Verfassern dieser Studie einen der grössten Datensätze mit realen menschlichen Äusserungen im Internet darstellt. Entsprechend den Ergebnissen dieser Untersuchung ist die Sprache im Allgemeinen positiv emotional geladen und signifikant anders als neutral. Den Autoren zufolge enthalten negative Wörter mehr Informationsgehalt als positive, weil sie seltener vorkommen.

[7] http://books.google.com/ngrams/datasets

Die Erkenntnis, dass positive Wörter häufiger geäussert werden, deckt sich mit der sozialpsychologischen Forschung, die besagt, dass soziale Bindungen durch positive Emotionen gestärkt werden. Dadurch erhalten Gemeinschaften mit positiv verzerrter Kommunikation einen evolutionären Vorteil und positive Sätze können zur sozialen Norm werden.

Rozin et al. (2010) führen als weiteren Grund für die vermehrte Verwendung positiver Wörter an, dass positive Ereignisse im Leben im Allgemeinen häufiger vorkommen als negative. Weiter werden in dieser Studie einige Eigenheiten bei der Verwendung positiver und negativer Wörter beschrieben:

- Negative Wörter werden oft aus einer positiven Wurzel und einem Präfix gebildet (z. B. *unhappy*, *unpleasant*), während dem das Gegenteil aussergewöhnlich ist (z. B. *unselfish*).
- Negierte positive Adjektive haben eine negative Valenz (z. B. *not good* ist negativ), negierte negative Adjektive hingegen sind tendenziell neutral (z. B. *unsad* ist neutral und nicht *happy*).
- In Konjunktionen (oder) wird typischerweise zuerst das positive Adjektiv vor dem gegenteiligen negativen Adjektiv verwendet, wobei Ausnahmen natürlich vorkommen.

2.5 User generated content

User generated content (*UGC*) wird von *Ochoa und Duval* (2008) als Inhalt definiert, zu dessen Erstellung ein gewisses Mass an kreativer Anstrengung nötig war, öffentlich über das Internet verfügbar ist und der nicht in professioneller Art und Weise hergestellt wurde. Diese Definition ist nicht unumstritten und trifft nicht immer zu, aber laut *Ochoa und Duval* (2008) gibt sie doch die Hauptcharakteristiken sehr unterschiedlicher Inhalte wieder. UGC ist unter anderem auf Blogs, Videoplattformen, Fotodiensten oder auch Wikipedia und OpenStreetMap zu finden.

Wenn die nutzergenerierten Daten explizite geografische Informationen enthalten (beispielsweise Koordinaten), spricht man oft von *volunteered geographic information* (VGI). Die geografischen Informationen werden entweder zielgerichtet gesammelt und aufbereitet, wie etwa bei OpenStreetMap, oder sie entstehen quasi als Nebenprodukt, wie dies bei den georeferenzierten Fotos von Flickr der Fall ist.

Ochoa und Duval (2008) haben das Nutzerverhalten bei der Erstellung von UGC analysiert. Es zeigte sich dabei, dass typischerweise viele Nutzer wenig beitragen, einige wenige aber extrem viel. Dieses Verhalten wurde von *Nielsen* (2006) durch die Faustregel 90-9-1 auf

den Punkt gebracht: 90 % der Nutzer beteiligen sich nicht an der Erstellung, 9 % ab und zu und 1 % sind sogenannte «heavy contributors».

Gemäss *Purves et al.* (2011) gehen die meisten Autoren bei der Analyse von *volunteered geographic information* davon aus, dass einzelne Nutzer aufgrund der grossen Nutzerzahl keine Verzerrung in den Daten verursachen können. *Hollenstein und Purves* (2012) haben aber in ihrer Arbeit festgestellt, dass dies nicht immer stimmt und einzelne Nutzer in einer bestimmten Region dominierend sein können.

Bei nutzergenerierten Daten stellt sich die Frage nach der Qualität und Glaubwürdigkeit in erhöhtem Mass, denn diese Daten stammen oftmals von Menschen ohne spezifischen akademischen Hintergrund des betreffenden Gebietes, wie *Tulloch* (2007) erwähnt. Ausserdem erfolgt die Datenerhebung typischerweise nicht nach wissenschaftlichen Kriterien (*Flanagin und Metzger* 2008).

Andererseits argumentieren *Hill und Ready-Campbell* (2011), dass durch die «Weisheit der Masse» gute Urteile gefällt werden können, weil diese aus unterschiedlichen Menschen mit unterschiedlichen Informationen besteht. So gehen diese Autoren davon aus, dass man beispielsweise durch die Bewertungen eines Youtube-Videos dessen Downloadzahl vorhersagen kann. Ebenso kann man davon ausgehen, dass der ScenicOrNot-Datensatz, der ebenfalls durch Nutzer erstellt wurde, nicht rein zufällige sondern sinnvolle Meinungen über die Landschaften enthält, wie auch die Arbeit von *Stadler* (2010) gezeigt hat.

3 Datengrundlagen

3.1 Geograph

Die Landschaftsbilder, auf denen die Untersuchungen der vorliegenden Arbeit beruhen, stammen vom Geograph-Projekt[1], dessen Ziel darin besteht, für jeden einzelnen Quadratkilometer Grossbritanniens, Irlands und der Isle of Man eines oder mehrere repräsentative Bilder der Landschaft zu sammeln und der Öffentlichkeit sowie der Forschung zur Verfügung zu stellen. Das verwendete Quadratkilometerraster wird durch das National Grid des Ordnance Survey Great Britain gebildet. Die Bilder werden von Freiwilligen gemacht, wobei als Ansporn im Sinn eines Spiels Punkte an diejenigen Fotografen vergeben werden, die als erste ein Foto einer Rasterzelle hochladen (*Geograph* 2012). Neben den eigentlichen Fotos sind die Fotografen angehalten, einen Titel zu vergeben und eine kurze Beschreibung über das Foto zu verfassen. Die Fotos sind explizit georeferenziert, das heisst, sie können geografisch exakt verortet werden. Die Fotos auf Geograph werden moderiert und können, falls sie ungeeignet sind, zurückgewiesen werden. Als ungeeignet werden beispielsweise Fotos genannt, die hauptsächlich Personen zeigen, weil dies nicht dem Zweck von Geograph entspricht. Die Gemeinschaft von Geograph wird von *Purves et al.* (2011) als «geography enthusiasts» beschrieben.

Die Metadaten zu den Fotos wurden anfangs September 2012 heruntergeladen. Die Fotos stammen von 11'496 Fotografen, die zu diesem Zeitpunkt insgesamt 3'087'160 Bilder hochgeladen hatten. Davon sind 2'650'123 (85,84 %) repräsentative Fotos und 437'037 sogenannte ergänzende Fotos. 157'023 Fotos wurden in Irland oder Nordirland gemacht. Diese sind nicht Teil der Untersuchung, da für diese keine Landschaftsbewertungen vorliegen. Die Tab. 3.1 gibt einen Überblick, wieviele Bilder wo aufgenommen wurden.

	Geographs	ergänzende Bilder	Total
Grossbritannien	2'519'767	410'370	2'930'137
Irland	130'356	26'667	157'023
Gesamt	2'650'123	437'037	3'087'160

Tabelle 3.1: Anzahl Geographs und ergänzende Bilder im Datensatz

[1] http://www.geograph.org.uk

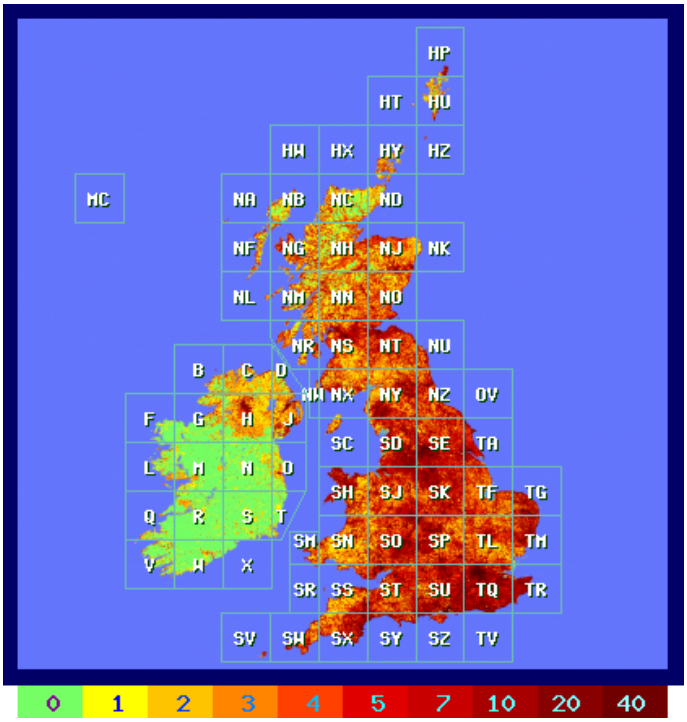


Abb. 3.1: Geografische Verteilung der Anzahl Geograph-Bilder auf den britischen Inseln
Quelle: http://www.geograph.org.uk/map/toVJ5oOXXJ0oX.VJFoOXXJfo-lNXJqo-NMJL5405oOh44V4lOuhVZuONu?depth=1, abgerufen am 7.12.2012

Die Abb. 3.1 zeigt, wie die Bilder auf Geograph geografisch verteilt sind. In Grossbritannien nimmt die Anzahl von Süden nach Norden tendenziell leicht ab. Schottland ist deswegen schlechter als die restlichen Gebiete erfasst. In der Republik Irland gibt es noch kaum Bilder, in Nordirland ist die etwas Anzahl grösser, diese beiden Regionen werden aber nicht weiter analysiert und die geringe Abdeckung ist deshalb nicht von Bedeutung.

3.1.1 Verteilung der Geograph-Fotografen

Die Landschaftsfotografien, die in ScenicOrNot bewertet wurden, stammen von insgesamt 4780 verschiedenen Fotografen. Ein grosser Teil dieser Fotografen hat nur wenige Bilder beigesteuert – einige wenige dafür aber sehr viele. Dies ist ein typisches Merkmal von nutzergenerierten Daten. In der Abb. 3.2 sind die Standorte der Fotos von Fotografen mit mehr als 1000 beigesteuerten Bildern dargestellt, die nicht nur in Geograph vorhanden sind, sondern auch von ScenicOrNot bewertet wurden. Diese Bedingung trifft auf 35 Fotografen zu, die zusammen 65'654 Bilder hochgeladen haben. Wie die Landschaftsbewertungen sind auch die Fotografen nicht zufällig im Raum verteilt. Das Gegenteil ist der Fall, denn die fleckenhafte Verteilung der Farben zeigt, dass viele dieser Fotografen

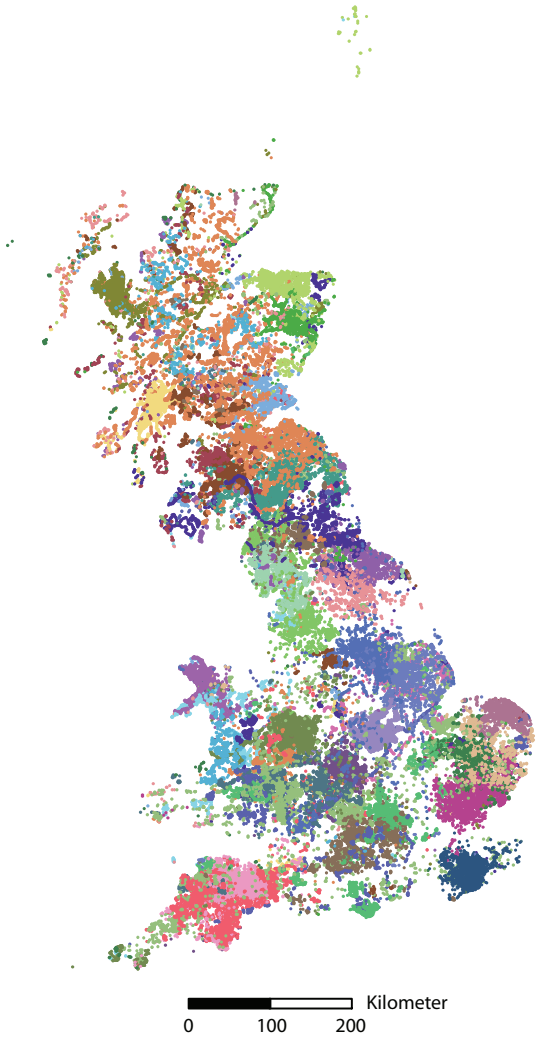


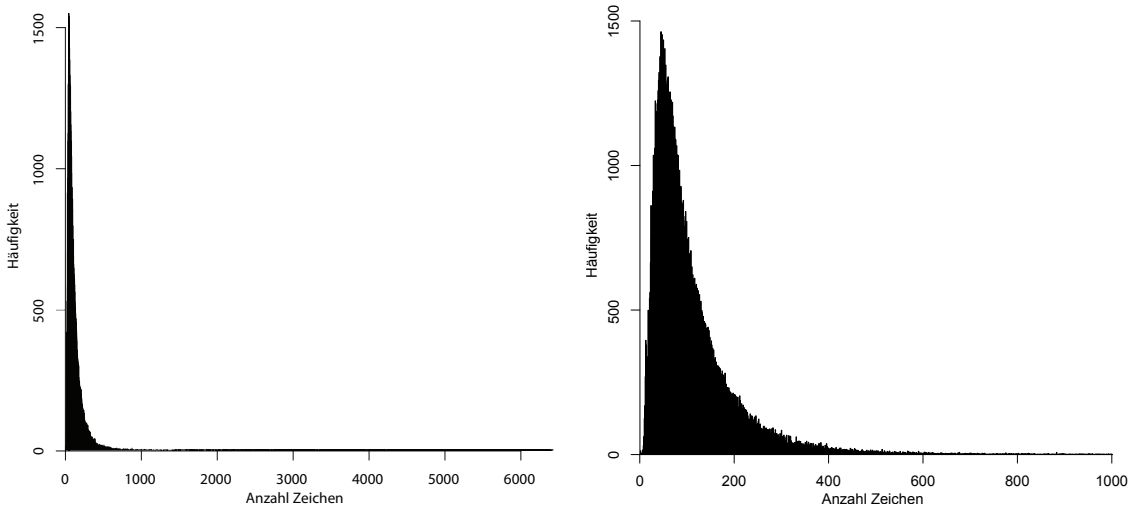
Abb. 3.2: Fotostandorte eingefärbt nach Geograph-Usern. Dargestellt sind die Standorte der Fotos von Freiwilligen, die mehr als 1000 bewertete Fotos beigesteuert haben.

ein zusammenhängendes Gebiet umfassend abgedeckt haben. Bei einigen Fotografen, vor allem in den nördlicheren Regionen, sind auch linienförmige Elemente erkennbar, die darauf hinweisen, dass die betreffenden Fotografen ihre Bilder während einer Reise geknipst haben, die nun in den Daten nachvollziehbar ist.

Bezüglich der Landschaftsbewertungen ist diese Clusterung ohne weitere Auswirkung, da diese von anderen Personen vorgenommen wurden. Aber weil die Landschaftsbewertungen, wie später im Abschnitt 3.2.1 auf Seite 30 gezeigt wird, räumlich autokorreliert sind, ist es denkbar, dass die Bewertung einzelner Wörter, die von einem Fotograf überdurchschnittlich oft verwendet wurden, durch diesen Fotografen verzerrt ist. Diesem Umstand wird durch die Erstellung von nutzergetrennten Lexika (siehe Abschnitt 5.4.1 auf Seite 43) Rechnung getragen.

3.1.2 Geograph-Kommentare

Das Geograph-Projekt beabsichtigt, alle Fotos mit einem Kommentar zu versehen. Trotz diesem Ziel ist das bei den Fotos, die von ScenicOrNot bewertet wurden, nicht immer der Fall. 53'561 der 211'380 Bilder verfügen über keinen Kommentar – das entspricht 25.3 %. Die Abb. 3.3 zeigt, wie die Kommentarlängen verteilt sind. Abb. 3.3 (a) zeigt das ganze Spektrum; Abb. 3.3 (b) einen Ausschnitt, der die Anzahl Zeichen von 1 bis 1000 umfasst. In dieser Bandbreite bewegen sich fast alle Kommentare, nur 223 Bilder haben noch längere Texte.



(a) Anzahl Zeichen: 1 bis Maximum

(b) Anzahl Zeichen: 1 bis 1000

Abb. 3.3: Histogramme der Länge der Geograph-Kommentare.

Wortarten

Tab. 3.2 auf der nächsten Seite zeigt die Anteile der wichtigsten Wortarten in den Geograph-Kommentaren. Dazu wurde für jeden Kommentar ein Part-of-Speech-Tagging durchgeführt und anschliessend die damit bestimmten Wortarten aufsummiert. Auffällig ist, wie gut der Anteil der Nomen (36.87 %) in den Kommentaren mit einem Wert aus der Literatur übereinstimmt, der den durchschnittlichen Anteil von Nomen in englischen Texten mit 37 % angibt (*Hudson* 1994). Diese Tatsache ist ein Indikator dafür, dass die Kommentare aus richtigen Texten bestehen und nicht beispielsweise nur aus Stichworten, weil es recht unwahrscheinlich wäre, dass die Nomenverteilung dann so exakt dem Durchschnitt englischer Texte entsprechen würde.

Im Umkehrschluss kann man auch, wenn man *a priori* davon ausgeht, dass die Kommentare durchschnittlichen englischen Texten entsprechen, folgern, dass der Part-of-Speech-Tagger ziemlich zuverlässig funktioniert.

Wortart	**Anteil**
Nomen	36.87 %
Verben	13.26 %
Adjektive	7.07 %
Adverben	4.46 %
Übrige	38.33 %

Tabelle 3.2: Anteile der wichtigsten Wortarten in den Geograph-Kommentaren

Einschränkungen dieser Datengrundlage, die sich aufgrund des wissenschaftlichen Hintergrundes ergeben, sind im Abschnitt 4.1.1 auf Seite 34 näher erläutert.

3.2 ScenicOrNot

mySociety[2] ist eine Organisation, die mithilfe von Websites die britische Gesellschaft demokratisch vernetzen will. Eines der Projekte ist ScenicOrNot[3], das für fast 95 % der britischen 1km-Rasterzellen ein Bild von geograph.org.uk zur Verfügung hat. Insgesamt befinden sich 217'000 Bilder in der Datenbank von ScenicOrNot, die allesamt aus Grossbritannien stammen. Für jede Rasterzelle wird nur ein Bild verwendet, auch wenn mehrere vorhanden wären. Es ist unklar, wie die Auswahl in Zellen mit mehreren Bilder vorgenommen wird. ScenicOrNot merkt an, man könne nichts dagegen tun, wenn ein Bild nicht repräsentativ ist, ausser dass neue Bilder hochgeladen werden können, sobald alle bereits vorhandenen Bilder mindestens drei Mal bewertet sind.

Das Ziel von ScenicOrNot ist, einen frei verfügbaren Datensatz der Landschaftsschönheit von ganz Grossbritannien zu erstellen (*ScenicOrNot* 2012). Um dieses Ziel zu erreichen, können beliebige Internetnutzer auf der Website von ScenicOrNot, die in Abb. 3.4 auf der nächsten Seite dargestellt ist, die Landschaftsbilder auf einer Skala von 1 (nicht schön, «not scenic») bis 10 (sehr schön, «very scenic») anonym bewerten. Die Bewertung geschieht durch Klicken einer Zahl auf der Skala im oberen Bereich der Website. Nach der Bewertung wird das Bild im linken Bereich zusammen mit seinem Titel und der durchschnittlichen Bewertung sowie der Stimmenzahl angezeigt. Bevor das Bild bewertet wird, werden weder Titel noch Beschreibung angezeigt, allerdings kann man durch Aufrufen des Links

[2] http://www.mysociety.org
[3] http://scenic.mysociety.org

Abb. 3.4: Website von ScenicOrNot zur Bewertung der Landschaften
Quelle: http://scenic.mysociety.org, abgerufen am 12.12.2012

Photo unterhalb des Bildes diese Informationen trotzdem erhalten. Trotz der einfachen Benutzeroberfläche kann nicht ausgeschlossen werden, dass die Bewertungsskala falsch verstanden wird und deshalb bei als unästhetisch empfundenen Landschaften die Bestnote 10 vergeben wird. Die Suche im Datensatz nach Bildern mit grosser Varianz zeigt, dass dies in Einzelfällen wohl auch vorgekommen ist. Das in Abb. 3.5 auf der nächsten Seite gezeigte Bild hat vier Mal die Note 1 und zwei Mal die Bestnote 10 erhalten und es ist bei Betrachtung des Bildes naheliegend, dass die Vergabe der Bestnote vermutlich irrtümlich erfolgt ist.

ScenicOrNot stellt diese Daten zum Download zur Verfügung, allerdings nur von Bildern, die mindestens drei Mal von unterschiedlichen Menschen bewertet wurden, um den Einfluss von Missbrauch und «schrecklichem Geschmack» (*ScenicOrNot* 2012) zu mindern. Zum Zeitpunkt der Datenbeschaffung anfangs September 2012 traf diese Bedingung bei 211'380 Bildern zu. Es wurden folglich noch nicht alle Rasterzellen, für die Bilder vorhanden sind, mindestens 3 Mal bewertet. Neben den Koordinaten und der durchschnittlichen Bewertung sind auch die Einzelstimmen, die Varianz und die Information, um welches Geograph-Bild es sich handelt, enthalten. Unter der Annahme, dass ein Bild nicht mehrmals von derselben Person bewertet worden ist, weiss man die Anzahl Bewerter pro Bild. Wie viele verschiedene Personen aber insgesamt Stimmen abgegeben haben, geht aus den Daten nicht hervor, weil nicht klar ist, welcher User welche Bilder bewertet hat.

Abb. 3.5: Beispiel eines Bildes mit seltsamer ScenicOrNot-Bewertung. Dieses Bild hat vier Mal die Note 1 und zwei Mal die Bestnote 10 erhalten. Es handelt sich um den Parkplatz eines Golfplatzes südlich von Bristol.
Quelle: http://www.geograph.org.uk/photo/494515, abgerufen am 10.12.2012

Das Histogramm in Abb. 3.6 auf der nächsten Seite zeigt, wie oft die bewerteten Bilder in den verschiedenen Bewertungsklassen vorkommen. Die Grafik zeigt eine rechtsschiefe Verteilung, was bedeutet, dass in den tieferen Bewertungsklassen mehr Bilder vorkommen als in den hohen. Insbesondere die schönste Klasse, Bilder mit einem Durchschnitt von 9 – 10, kommt nur sehr selten vor. Die grössere Zahl unterdurchschnittlich bewerteter Bilder bedeutet aber nicht notwendigerweise, dass die ästhetisch wenig gefälligen Landschaften überwiegen. Es ist denkbar, dass urbane, tendenziell schlechter bewertete Gebiete häufiger fotografiert wurden, weil sie einfacher zu erreichen sind als etwa die schottischen Highlands, die, wie man auf der Karte 3.7 sieht, gut bewertet sind. Eine an sich schöne Landschaft kann auch schlecht bewertet worden sein, weil die Fotografie davon als qualitativ schlecht betrachtet wird, was eine schlechte Bewertung nach sich ziehen kann. Eine qualitativ schlechte Aufnahme einer unschönen Landschaft landet aber ebenso in den tief bewerteten Kategorien. Ferner gibt es möglicherweise einen grösseren Konsens darüber, was nicht schön ist, während die Meinungen über schöne Landschaften stärker differieren.

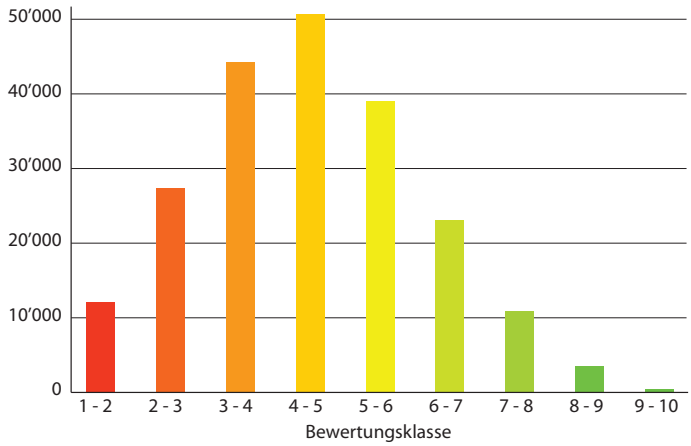


Abb. 3.6: Histogramm der ScenicOrNot-Bewertungen (eigene Darstellung)

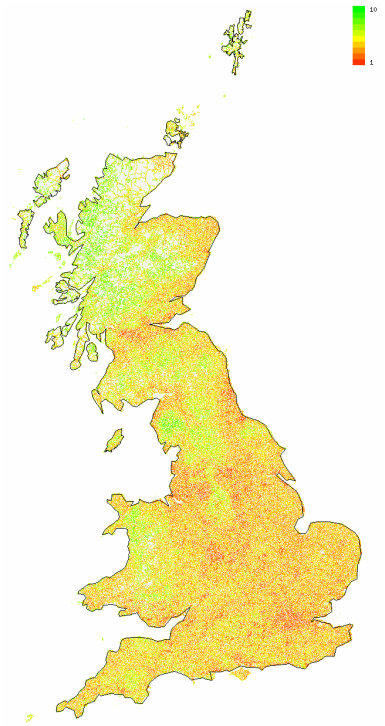


Abb. 3.7: Übersicht der ScenicOrNot-Bewertungen in Grossbritannien – die Bewertung nimmt im Allgemeinen in nördlicher Richtung zu. Jedes Pixel stellt einen Quadratkilometer dar (eigene Darstellung).

Die Abb. 3.7 gibt eine Übersicht der Verteilung der durchschnittlichen Bewertungen von ScenicOrNot in Grossbritannien. Ausserdem ist auch ersichtlich, in welchen Gebieten weniger Landschaftsbewertungen vorliegen. Diese sind hauptsächlich in Schottland zu finden, weil dort auch die Geograph-Abdeckung weniger dicht ist (siehe auch Abb. 3.1 auf Seite 23).

Die Abb. 3.8 zeigt einen Scatterplot der durchschnittlichen Bewertungen und ihrer Varianzen. Die Varianz zeigt die Einigkeit der Bewertenden in der Stimmenverteilung. Der Grossteil der Varianzen ist kleiner als 5, wie die Grafik zeigt. Berechnungen ergeben, dass 50 % aller Bilder von ScenicOrNot eine Varianz von weniger als 2.4 aufweisen. Die glockenförmige Form zeigt ausserdem, dass die Varianzen bei den mittleren Bewertungen am grössten sind und an den Extremen abnehmen.

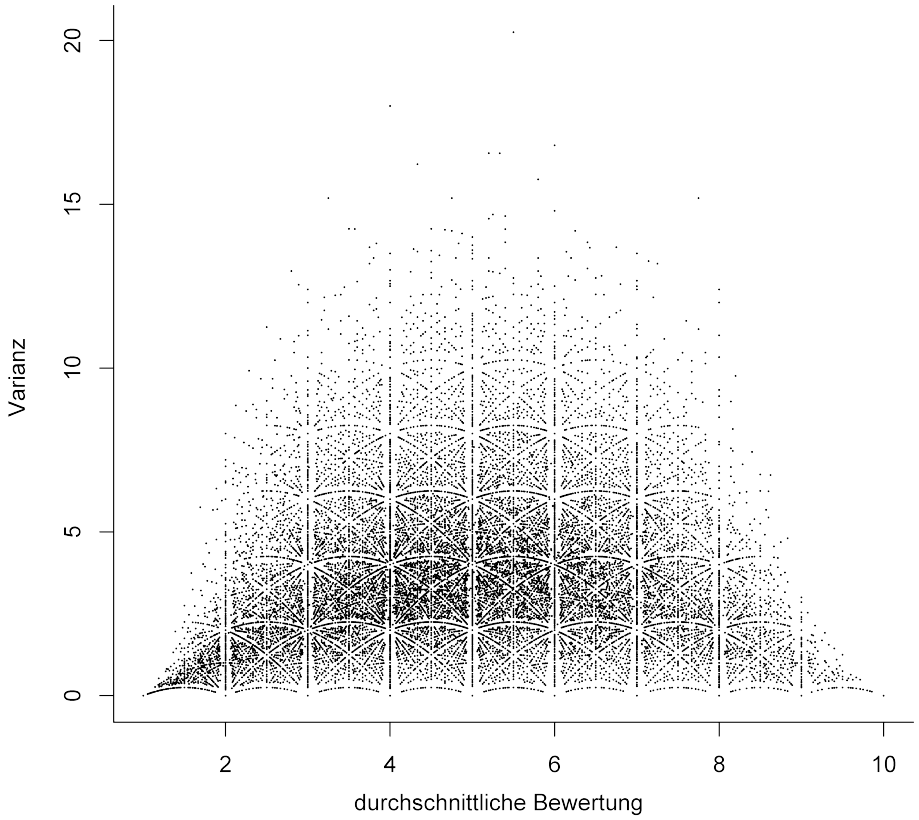


Abb. 3.8: Scatterplot der Bewertung vs. Varianz der ScenicOrNot-Bewertungen (eigene Darstellung).

3.2.1 Räumliche Autokorrelation

Die Abb. 3.7 auf der vorherigen Seite zeigt, dass die Landschaftsbewertungen nicht zufällig im Raum verteilt sind (dann würde es einem Rauschen gleichen), sondern einem bestimmten Muster folgen. Es kann deshalb angenommen werden, dass die Landschaftsbewertungen räumlich autokorreliert sind, also ähnliche Werte geclustert vorkommen. Diese Annahme ist in Übereinstimmung mit Tobler's erstem Gesetz der Geografie, das besagt, dass alles miteinander in Verbindung steht, nähere Dinge sich aber ähnlicher sind als weiter entfernte (*Tobler* 1970). Um die räumliche Autokorrelation zu quantifizieren, kann

Moran's I berechnet werden, ein Wert, der bei völlig positiver Autokorrelation 1 und bei völlig negativer Autokorrelation (ein Schachbrettmuster) –1 ergibt.

Die Anzahl der Landschaftsbewertungen, die in ScenicOrNot vorhanden ist, überfordert gängige Programme zur Berechnung von *Moran's I*. Um diesem Problem zu begegnen, wurde die Datenmenge durch ein Resampling reduziert. Dazu wurden die Landschaftsbewertungen in einem neuen Raster mit einer Zellgrösse von 7,5 km respektive 10 km zusammengefasst. Von allen ursprünglichen Werten, die in eine Zelle geflossen sind, wurde der durchschnittliche Wert berechnet und dieser der Zelle zugewiesen. Diese Raster sind in der Abb. 3.9 dargestellt. Die Berechnung von *Moran's I* mit Hilfe einer räumlichen Gewichtsmatrix ergibt einen Wert von 0,617 (7,5 km Zellgrösse) resp. 0,628 (10 km Zellgrösse). Die Wahrscheinlichkeit, dass diese Werte zufällig zustande gekommen sind, ist kleiner als 1 %. Die Vermutung, die Landschaftsbewertungen seien räumlich autokorreliert, wird durch die Berechnungen also gestützt. Es gilt aber zu beachten, dass sich die räumliche Autokorrelation durch die Vergrösserung der Zellgrössen und der damit verbundenen Glättung der Werte verstärkt hat, wie dies auch in den beiden dargestellten Zellgrössen zu beobachten ist. Man kann deshalb davon ausgehen, dass die räumliche Autokorrelation der ursprünglichen Daten geringer, aber dennoch vorhanden ist.

Einschränkungen dieser Datenquelle, die sich aufgrund des wissenschaftlichen Hintergrundes ergeben, sind im Abschnitt 4.1.1 auf Seite 35 beschrieben.

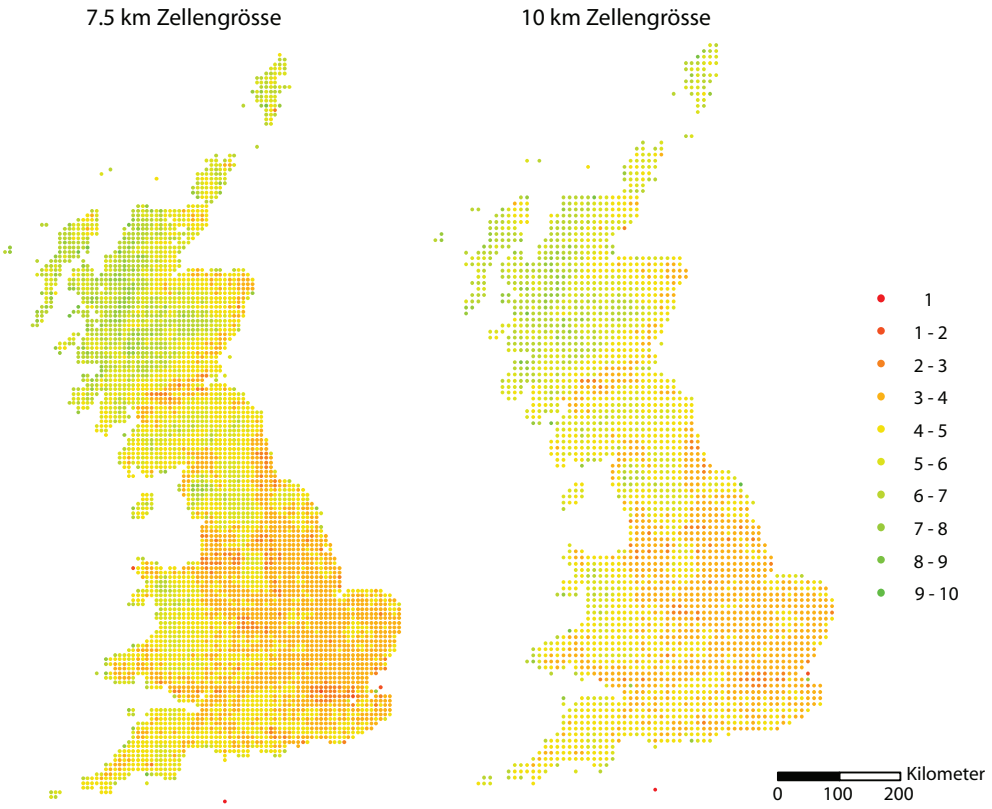


Abb. 3.9: Raster der Landschaftsbewertungen mit 7,5 resp. 10 km Maschenweite zur Berechnung von *Moran's I*.

3.3 SentiWordNet

SentiWordNet ist ein Lexikon, das basierend auf den 117'659 Synsets von WordNet für alle diese Einträge einen Stimmungswert (positiv oder negativ zwischen 0 und 1) angibt. Ein Synset ist eine Gruppe von Worten, die die gleiche inhaltliche Bedeutung haben. Ein Synset kann auch teils positiv und teils negativ sein. Zusammen mit einem dritten Wert, nämlich der Objektivität, beträgt die Summe des Gesamtwerts immer 1. Auch ein Objektivitätswert von 1 ist möglich, wenn ein Wort keine Stimmung übermittelt. SentiWordNet wurde durch eine semi-überwachte Lernphase und einen anschliessenden iterativen Prozess kreiert, wobei der iterative Prozess beendet wurde, nachdem die Resultate konvergierten. SentiWordNet wurde zuerst in Version 1.0 veröffentlicht und später auf Version 3.0 aktualisiert. Diese Version wurde in dieser Arbeit verwendet und ist im Internet zum Download[4] verfügbar. SentiWordNet wird von *Baccianella et al.* (2010) in ihrem Paper beschrieben.

[4]http://sentiwordnet.isti.cnr.it/

Ein alternatives Wörterbuch ist der «Maryland dictionary» (*Mohammad et al.* 2009), der aber im Gegensatz zu SentiWordNet keine Informationen über die Wortarten (*part of speech*) enthält.

An grossen Wörterbüchern wie SentiWordNet wird kritisiert, dass durch eine grössere Anzahl Wörter auch mehr *noise*, also Rauschen, entsteht (*Taboada et al.* 2011).

3.4 Ortsverzeichnis

Für die Toponymerkennung wird das Ortsverzeichnis[5] von Ordnance Survey (Landesvermessung Grossbritannien) als Datengrundlage verwendet. In diesem auch als Gazetteer bezeichneten Datensatz sind 258'487 Ortsbezeichnungen Grossbritanniens mit ihren jeweiligen Koordinaten vermerkt. Der Datensatz ist auf der Website von Ordnance Survey frei verfügbar. Er wird als der detaillierteste verfügbare Datensatz beschrieben.

3.5 Synonym-Wörterbuch

Die Synonym-Suche, die in dieser Arbeit angewendet wird, basiert auf dem frei zugänglichen *Moby Thesaurus* des «Projekt Gutenberg»[6] von Grady Ward. Im Thesaurus sind Synonyme der englischen Sprache verzeichnet. Weil auch die anderen Daten im MySQL-Format vorliegen und dieses als Grundlage für die Berechnungen dient, wurde der Thesaurus in diesem Format in die Arbeit integriert. Die MySQL-Version ist ebenfalls im Internet verfügbar[7]. Diese ermöglicht eine rasche und strukturierte Suche nach Synonymen. Der Thesaurus beinhaltet Synonyme für gut 100'000 englische Wörter.

[5] http://www.ordnancesurvey.co.uk/oswebsite/products/50k-gazetteer/index.html
[6] http://www.gutenberg.org/ebooks/3202
[7] http://code.google.com/p/moby-thesaurus/downloads/list

4 Implikationen

4.1 Implikationen des wissenschaftlichen Hintergrunds für die Arbeit

Nachfolgend wird erörtert, welche Implikationen die Erkenntnisse, die sich aus dem wissenschaftlichen Hintergrund der einzelnen Gebiete ergeben, auf die Arbeit haben. Am Schluss des Kapitels werden die Forschungslücken beschrieben.

4.1.1 Landschaftsevaluierung

Fotografien

Die Fotografien, die als Grundlage für die Landschaftsbewertungen von ScenicOrNot dienen, wurden von Freiwilligen im Rahmen des Non-Profit-Projektes geograph.org.uk erstellt. Die Auswahl der Fotostandpunkte bei diesem Projekt genügt nicht den wissenschaftlichen Anforderungen an Samplingstrategien, die von *Hull et al.* (1989) beschrieben wurden (siehe Abschnitt 2.2.3 auf Seite 9). Sie können aber als Anwendung der von *Hull et al.* (1989) ebenfalls beschriebenen «participant photography» betrachtet werden. Deren Vorteil ist, dass die abgelichteten Szenen den Laienfotografen als wichtig für die Landschaft erscheinen und dass typischerweise Fotostandpunkte ausgewählt werden, die von der Öffentlichkeit häufig besucht werden. Ausserdem wird auf geograph.org.uk gefordert, dass die fotografierten Szenen klar eines der hauptsächlichen geografischen Merkmale des Quadratkilometers zeigen müssen[1] und weiter wird auch erwähnt, dass zur repräsentativen Darstellung eines Quadratkilometers unter Umständen mehrere Fotos gemacht werden müssen, weil oft mehr als ein Landschaftstyp vorkommt[2]. Aufgrund der grossen Menge von Bildern und den Vorteilen der «participant photography» kann davon ausgegangen werden, dass die britischen Landschaften repräsentativ erfasst wurden. Trotzdem ist festzuhalten, dass die Auswahl des Fotostandpunktes einen grossen Einfluss auf die spätere Bewertung haben kann.

Es gibt keine Vorgaben zu den Witterungsbedingungen, dem Tageszeitpunkt oder der Saison, die beim Fotografien berücksichtigt werden sollten und die einen Einfluss auf

[1] FAQ von geograph.org.uk: http://www.geograph.org.uk/faq3.php#55 abgerufen am 5.12.2012
[2] Quickstart Guide von geograph.org.uk: http://www.geograph.org.uk/article/Geograph-Quickstart-Guide abgerufen am 5.12.2012

die wahrgenommene Landschaftsqualität haben können (*Hull und McCarthy* 1988, siehe Abschnitt 2.2.4 auf Seite 10). Dafür hat geograph.org.uk das Konzept der «ergänzenden Bilder» (supplemental image), eine Kategorie, in die Bilder eingeteilt werden, die nicht die Bedingungen der landschaftlichen Repräsentativität erfüllen, aber trotzdem einen Eindruck des betreffenden Quadratkilometers vermitteln. Ein Beispiel eines ergänzenden Bildes ist in Abb. 4.1 gezeigt, das im Übrigen in ScenicOrNot nicht bewertet wurde. Die ergänzenden Bilder stellen mit 420 Fotos nur 0.19 % der in ScenicOrNot bewerteten Bilder dar.

Abb. 4.1: Beispiel eines ergänzenden Bildes von geograph.org.uk: Sonnenaufgang bei Portree (Isles of Skye)
Quelle: http://www.geograph.org.uk/photo/2719508, abgerufen am 5.12.2012

Bewertung

Die Landschaftsbewertungen von ScenicOrNot beruhen auf Fotografien mit den oben beschriebenen Einschränkungen. Gemäss *Stamps* (1990) sind Fotos als Ersatz für Feldbegehungen zulässig (siehe S. 6). Die ScenicOrNot-Bewertungen werden im Internet vergeben, was ebenfalls zulässig ist (siehe Abschnitt 2.2.2 auf Seite 8). Allerdings gibt es keine Informationen über die bewertenden Personen, da die Evaluierung komplett anonym geschieht. Auch das Umfeld (z. B. Lärm) und der Zeitpunkt können einen Einfluss auf die Bewertung haben (siehe Abschnitt 2.2.4 auf Seite 10), aber auch über diese Faktoren gibt es keine Kontrolle, was bei anderen Untersuchungen normalerweise der Fall ist. Weiter

ist davon auszugehen, dass sich tendenziell eher an Landschaften interessierte Personen an der Bewertung beteiligen, da die Teilnahme auf Freiwilligkeit basiert. Man kann also annehmen, dass Selbstselektion vorliegt. Dies senkt aber gleichzeitig die Wahrscheinlichkeit für «Vandalenbewertungen», was vorteilhaft ist. Durch das einfache Bewertungsschema von ScenicOrNot – ein Wert auf einer Skala von 1-10 – gibt es keinerlei Information darüber, wie die Landschaftsbewertung zustande gekommen ist, was aber gemäss *Arthur et al.* (1977) wünschenswert wäre. Es kann auch nicht ausgeschlossen werden, dass bloss die Qualität der Fotografie und nicht die Landschaft selbst bewertet wurde.

Wie in Abschnitt 2.2.5 auf Seite 10 beschrieben wurde, kann eine Landschaftsbewertung durch die Beschriftung eines Bildes mit Begriffen, die entweder menschlichen Einfluss oder Naturnähe implizieren, beeinflusst werden (*Hodgson und Thayer* 1980). Diese Gefahr besteht bei den ScenicOrNot-Bewertungen aber nicht, da zum Zeitpunkt der Bewertung weder der Titel noch die Beschreibung des Fotos angezeigt werden.

Trotz den genannten Einschränkungen ist ScenicOrNot ein sinnvoller Datensatz, wie auch die Masterarbeit von *Stadler* (2010) gezeigt hat. *Stadler* (2010) konnte über 60 % der Varianz der Schönheitsbewertungen von ScenicOrNot durch die Landbedeckung erklären und konnte Ergebnisse, die in der Literatur zu finden sind (etwa die überdurchschnittliche Bewertung von Wasserflächen), bestätigen.

4.1.2 Opinion Mining und Meinungsäusserung

Eine verbreitete Aufgabenstellung in Opinion-Mining-Anwendungen ist die Klassifikation in objektive und subjektive Sätze (*Wiebe et al.* 2005), weil man Meinungen von Fakten trennen will. Ein möglicher Ansatz dazu ist die Verwendung der Ergebnisse von *Pak und Paroubek* (2010) (siehe Abschnitt 2.3.2 und Abb. 2.2 auf Seite 16). Bei der vorliegenden Arbeit stellt sich aber die Frage, ob dieser Schritt nötig ist, denn es ist denkbar, dass Meinungen über Landschaften auch in üblicherweise objektiven Wörtern stecken (etwa *town* oder *forest*).

Auch die sogenannte *feature extraction* wird in der Literatur oft erwähnt (z. B. *Popescu und Etzioni* 2005), allerdings in den typscherweise behandelten Themen wie etwa Produktereviews, wo die Meinung über einzelne Gerätekomponenten von Interesse ist. Es ist durchaus möglich, dass sich die Begleittexte von Geograph nicht nur auf die Landschaftsästhetik beziehen, sondern beispielsweise auf die Geschichte eines Ortes oder dergleichen. Bei der Ermittlung der Stimmung gegenüber der Landschaft könnte es deshalb hilfreich sein, nur diejenigen Teile zu berücksichtigen, die die Landschaft und deren Ästhetik betreffen.

Die in Abschnitt 2.4 auf Seite 19 erwähnte *Pollyanna-Hypothese* legt nahe, dass positive Wörter häufiger verwendet werden. Es stellt sich deshalb die Frage, ob negative Wörter beim Opinion Mining stärker gewichtet werden sollen, zumal diese gemäss *Garcia et al.* (2011) einen höheren Informationsgehalt haben. Die Autoren selbst erwähnen zwar, dass dieser Sachverhalt berücksichtigt werden sollte, lassen aber offen wie. *Taboada et al.* (2011) lösen dieses Problem, indem sie den Wert einer negativen Äusserung nach Anwendung aller Modifikatoren um 50% erhöhen. Ein weiterer Hinweis auf die Stärke der ausgedrückten Emotion ist die Häufigkeit eines Wortes relativ zum gesamten Korpus. Je häufiger ein Wort vorkommt, desto geringer ist seine Emotionsstärke (*Kisilevich et al.* 2010). *Taboada et al.* (2011) gehen ebenfalls auf dieses Phänomen ein, und zwar indem sie dem n-ten Auftreten desselben Wortes innerhalb eines Textes vom ursprünglichen Stimmungswert nur den 1/n-ten Teil nehmen.

Bei den Begleittexten von Geograph handelt es sich um informelle Texte, die deshalb eine informelle Sprache mit entsprechenden Wörtern aufweisen. Wie *Taboada et al.* (2011) anmerken, sollte ein Opinion-Mining-Lexikon mit Stimmungswerten für Wörter auch informelle Sprache berücksichtigen. Einige Stichproben mit typischen informellen Wörtern zeigen, dass zumindest ein Teil dieser informellen Wörter im Opinion-Mining-Wörterbuch SentiWordNet enthalten ist. Es ist auch anzumerken, dass die Begleittexte nicht die Struktur von informellen Texten wie etwa bei MySpace aufweisen (vgl. Abschnitt 2.3.2 auf Seite 15). Trotzdem ist es wahrscheinlich, dass Rechtschreibfehler vorkommen.

4.1.3 User generated content

Die beiden verwendeten Datenquellen Geograph und ScenicOrNot gehören in die Kategorie *User generated content*, oder präziser, in die Kategorie *Volunteered Geographic Information* (*VGI*), da beide Datensätze nur durch die Einbindung des Internets entstehen konnten. Das im Abschnitt 2.5 auf Seite 20 geschilderte Verhalten der Nutzer trifft auch auf Geograph zu: Die Fotos stammen von insgesamt 11'496 Nutzern, von denen aber 2936 nur ein einziges beigesteuert haben, 425 Nutzer dafür jeweils über 1000. Über die Nutzerzahlen von ScenicOrNot und deren Verhalten ist nichts bekannt, aber unter Berücksichtigung der Faustregel von *Nielsen* (2006) kann man davon ausgehen, dass auch bei ScenicOrNot einzelne Nutzer überproportional viele Landschaftsbewertungen beigesteuert haben und dass deshalb eine gewisse Verzerrung zugunsten deren Vorlieben vorliegen könnte.

Bei den Geograph-Bildern und deren Begleittexten ist das von *Hollenstein und Purves* (2012) festgestellte Phänomen der Dominanz einzelner Nutzer in einer Region nicht auszuschliessen. Es ist vorstellbar, dass einzelne Nutzer die Region, aus der sie stammen, besonders umfassend auf Geograph dokumentiert haben. An Orten wo dies der Fall ist,

widerspiegelt eine Analyse der Begleittexte dann in erster Linie die Sichtweise dieser Person. Um solche Effekte zu minimieren, wäre es deshalb wünschenswert, dass benachbarte Quadratkilometerzellen von jeweils unterschiedlichen Fotografen besucht würden.

Über die Genauigkeit der erfassten Koordinaten der Fotos auf Geograph ist wenig bekannt, denn diese können entweder nachträglich mit Hilfe einer digitalen Karte oder bereits im Feld mit einem GPS-Empfänger georeferenziert werden. Bei beiden Vorgehensweisen sind Ungenauigkeiten unumgänglich. Für den Verwendungszweck der Daten in dieser Arbeit sind solche Ungenauigkeiten aber tragbar, denn die Aussagen über Landschaftsqualität beziehen sich auf grössere Gebiete und man kann davon ausgehen, dass die Zuordnung der Fotos zu Rasterzellen grösstenteils korrekt ist. Die Koordinaten werden nur zur Filterung der Toponyme verwendet.

4.2 Forschungslücken

Die meisten bisherigen Ansätze zur Landschaftsevaluierung beruhen, egal ob quantitative oder qualitative Methoden verwendet werden, auf Befragungen einer begrenzten Anzahl Probanden oder einzelner Experten sowie einer eher geringen Anzahl Bilder, wenn nicht mit Feldbegehungen gearbeitet wurde. Auch die Landschaftsbewertungen auf Feldbegehungen sind wegen dem logistischen Aufwand in der Zahl stark eingeschränkt. Diese Vorgehensweise hat Vorteile bezüglich der Kontrolle der Bedingungen, aber die Teilnehmerzahl und die geografische Ausdehnung ist typischerweise begrenzt. Auch *Roth* (2006), der Landschaftsbewertungen im Internet erhoben hat, hatte nur 321 Teilnehmer, die mindestens ein Foto evaluiert haben.

Die Daten von Geograph und ScenicOrNot hingegen ermöglichen die Analyse von einem grossen Gebiet (in diesem Fall Grossbritannien) mit detaillierter Auflösung und die Auswertung von hunderttausenden von Bewertungen, die durch Schwarmintelligenz entstanden sind. Mit Ausnahme der Arbeit von *Stadler* (2010), der dieselben, damals noch etwas kleineren Datensätze benutzt hat, weiss ich von keiner Arbeit, die so grosse Datensätze zur Landschaftsbewertung verwendet hat oder die computerlinguistische Methoden mit Landschaftsevaluierung kombiniert hat.

Der Zweck dieser Arbeit ist deshalb zu erforschen, wie die Analyse von Begleittexten von Landschaftsfotos einen Beitrag zur Landschaftsevaluierung leisten kann und wie sich eine Landschaftsbewertung mit computerlinguistischen Methoden schätzen lässt.

5 Methodisches Vorgehen

5.1 Vorbemerkungen

In der Abb. 5.1 ist eine Übersicht des Vorgehens von den verschiedenen Datenquellen zu den Ergebnissen dargestellt.

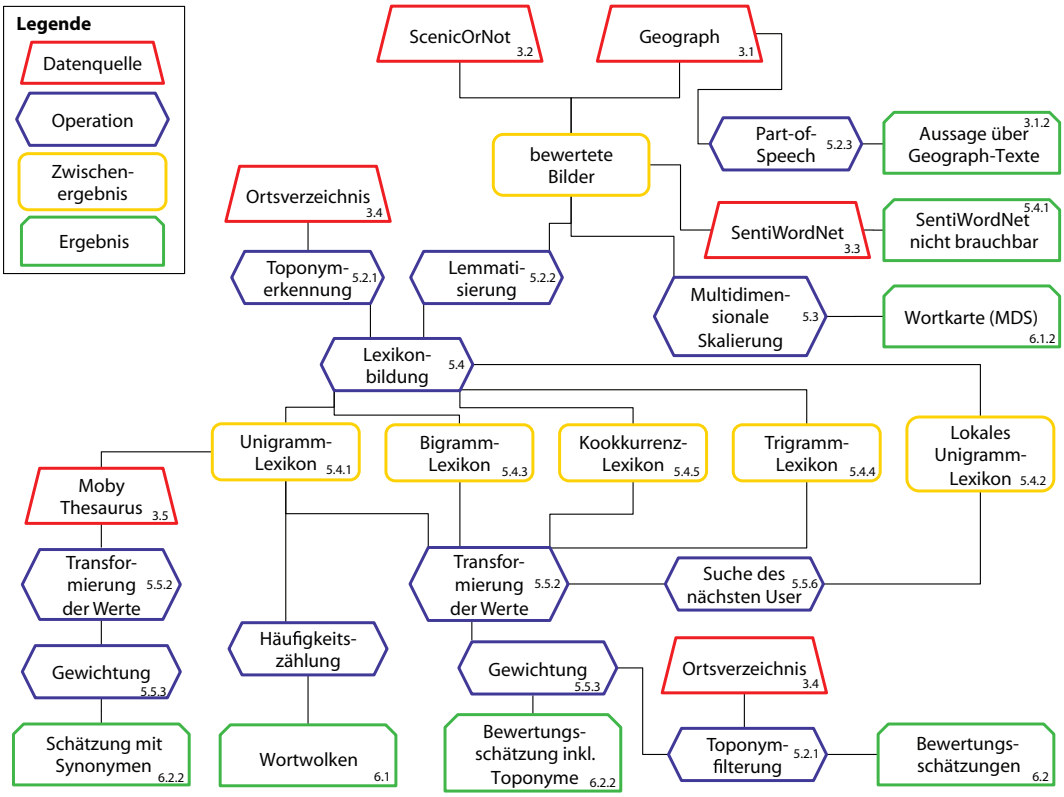


Abb. 5.1: Flowchart des Vorgehens. Die Zahl gibt den Abschnitt an, in dem die entsprechenden Ausführungen zu finden sind (eigene Darstellung).

5.1.1 Software und Datenstrukturen

Zur Auswertung der Daten und zur Schätzung der Landschaftsqualität wurde die Programmiersprache PHP verwendet, weil diese einen einfachen Zugriff auf die Datenbanken mit den Daten von Geograph, die im MySQL-Format verfügbar sind, und den anderen verwendeten Daten erlaubt. Sämtliche zusätzlich erstellten Tabellen (insbesondere die

verschiedenen Lexika) wurden ebenfalls in MySQL erstellt, weil dadurch strukturierte, computergenerierte Abfragen einfach zu implementieren sind.

Ferner wurden das Open-Source-Statistikpaket *R*, das GIS-Paket *ArcGIS* und zur Darstellung der Ergebnisse das Vektorzeichnungsprogramm *Adobe Illustrator* eingesetzt.

5.2 Werkzeuge

5.2.1 Toponymerkennung

Da Toponyme (Ortsnamen) keine Information über die Stimmung eines Textes geben, ist es angebracht, diese aus den Beschreibungen herauszufiltern und für die nachfolgenden Berechnungen nicht zu berücksichtigen. Dazu kann ein Gazetteer[1] verwendet werden, in dem die Toponyme mit ihren Koordinaten abgespeichert sind. Der verwendete Gazetteer von Ordnance Survey beinhaltet 258'487 Ortsnamen. Indem ein Skript alle Beschreibungen von Geograph Wort für Wort analysiert, können durch einen Abgleich mit der Toponym-Datenbank Toponyme erkannt werden, auch solche, die aus mehreren Wörtern bestehen. Um zu verhindern, dass Wörter, die zwar in einigen Fällen ein Toponym darstellen, aber nicht als solches verwendet wurden (beispielsweise «Forest») nicht fälschlicherweise immer als Toponym klassiert werden, kann der Aufnahmeort des analysierten Fotos als zusätzliche Information verwendet werden. Konkret bedeutet dies, dass die Distanz zwischen den im Gazetteer vermerkten Koordinaten des Toponyms und des Fotostandpunktes berechnet wird. Falls diese unter einem Grenzwert liegt, wird das Wort bzw. die Wortgruppe als Toponym klassiert. In dieser Arbeit wurden drei Grenzwerte verwendet, nämlich 5, 10 und 15 km. Innerhalb des grössten Radius werden naturgemäss mehr Begriffe als Toponym gekennzeichnet. Der grosse Radius von 15 km wird insbesondere verwendet, um in Sätzen wie «in the distance you see *xyz*», die typischerweise auf Aussichtspunkten vorkommen, auch Toponyme zu erkennen. Es gibt aber auch zahlreiche Beschreibungen im Stil von «an der Strasse zwischen *a* und *b*». Auch hier ist der grosse Grenzwert vorteilhaft, um die weiter entfernten Toponyme zu erkennen.

Mit diesem Vorgehen wird eine grosse Zahl von Toponymen gefunden – in jedem Lexikon jeweils gut 30'000. Trotzdem gibt es Begriffe, die fälschlicherweise nicht als Toponym erkannt wurden, etwa Berge, die in weiter Ferne am Horizont sichtbar sind und im Bildkommentar erwähnt werden. Allerdings gibt es auch eine Reihe falscher Positiver, welche in Wahrheit keine Toponyme darstellen. Um mit diesem Problem umzugehen, kann beim Ausschluss von Toponymen der Quotient zwischen der Gesamtzahl der gefundenen

[1] Ortsverzeichnis von http://www.ordnancesurvey.co.uk/oswebsite/products/50k-gazetteer/index.html

Wörter und der Anzahl der als Toponym markierten berechnet werden. Wenn dieser Wert im tiefen einstelligen Prozentbereich liegt, kann der entsprechende Begriff als normales Wort behandelt werden. *Hill* wird beispielsweise in jedem Lexikon über 4000 mal gefunden, aber nur ein- bis zweimal als Toponym klassiert. Deshalb würde es wenig Sinn machen, dieses Wort immer als Toponym zu betrachten. Stattdessen wird es als naturbezogener Begriff behandelt.

5.2.2 Lemmatisierung

Weil die Geograph-Kommentare aus natürlichem Text bestehen und nicht aus Tags wie etwa bei Flickr, kommen viele Wörter in unterschiedlichen Formen (Mehrzahl, konjugiert etc.) vor. Um die Wörter miteinander vergleichen zu können, ist es deshalb sinnvoll, die Grundform zu betrachten. Einerseits kann dafür *Stemming* verwendet werden, dessen Ziel darin besteht, den Wortstamm eines Wortes zu ermitteln. Allerdings entstehen dabei keine richtigen Wörter, sondern lediglich Wortstämme, was in dieser Arbeit nicht erwünscht ist. Die Alternative dazu ist die Verwendung eines *Lemmatisierers*, der die Grundform eines gewünschten Wortes zurückgibt (aus «is» wird z. B. «be»). In dieser Arbeit wird dazu die Open-Source-Implementierung phpMorphy[2] eingesetzt, die in der Programmiersprache PHP geschrieben ist und unter anderem auch für Englisch optimiert ist.

5.2.3 Part-of-Speech-Tagging

Part-of-Speech-Tagging bezeichnet die Bestimmung der in einem Satz vorkommenden Wortarten, also z. B. Nomen, Adjektive etc. Auch für diese Aufgabe wird in der vorliegenden Arbeit ein frei verfügbarer Algorithmus[3] verwendet, der ebenfalls in PHP geschrieben ist, was die Integration erleichtert.

5.3 Multidimensionale Skalierung

Die multidimensionale Skalierung ist ein Verfahren, mit dem die Ähnlichkeit bzw. Nähe von Objekten dargestellt werden kann. Diese Methode kann auch auf die Kommentare von Geograph angewendet werden, um damit die Nähe von bestimmten Begriffen zu analysieren. Es können aber nicht alle in den Daten vorkommenden Begriffe gleichzeitig in die Auswertung einfliessen, weil die Resultate dann unübersichtlich und nicht interpretierbar wären. Stattdessen muss eine Auswahl getroffen werden – der Fokus wurde dabei auf eine

[2] http://phpmorphy.sourceforge.net/dokuwiki/download
[3] http://phpir.com/part-of-speech-tagging

grosse Spannweite der Bewertungen und eine möglichst geringe Varianz gelegt. Mit diesen Vorgaben wurden 200 Begriffe ausgewählt. Um die Distanz der Wörter zueinander zu bestimmen, wird gezählt, wie oft die Begriffe zusammen in einem Kommentar vorkommen. Die Idee dahinter ist, dass sich zwei Begriffe inhaltlich näher sind, je öfter sie zusammen in einem Kommentar vorkommen.

Theoretisch ergeben sich mit den 200 Begriffen 40'000 Kombinationen, in denen sie auftreten können, allerdings sind darin auch die Begriffe mit sich selbst enthalten, deren Distanz per Definition null ist. Ausserdem wird jeweils Begriff A mit Begriff B und gleichzeitig B mit A kombiniert, so dass schlussendlich alle Kombinationen doppelt vorliegen. Es gibt deshalb insgesamt 19'900 eindeutige Kombinationen. Nach der Transformierung der paarweisen Vorkommen in eine symmetrische, normalisierte Distanzmatrix, die Distanzen zwischen null und eins enthält, kann mit einem Statistikprogramm für jeden Begriff eine Position in einem zweidimensionalen Raum berechnet werden. Die Positionen bilden, so gut es geht, die Distanzen der Distanzmatrix ab. Da für jeden Begriff die durchschnittliche Bewertung bekannt ist, kann diese Information ebenfalls dargestellt werden. Die Bewertung ist zu Beginn aber nur punktweise bekannt, nämlich an der Stelle des Begriffs im zweidimensionalen Raum. Um aus diesen Punktinformationen eine kontinuierliche Fläche darzustellen, müssen die Punktwerte über den Raum interpoliert werden. Dazu wurde in diesem Fall eine *Inverse Distance Weighting*-Interpolation (IDW) gewählt, deren Vorteil darin besteht, dass die interpolierte Fläche sich innerhalb des Minimums und des Maximums der Punktwerte bewegt. Nach der Interpolation können zusätzlich Isolinien berechnet werden, die Stellen mit gleicher Bewertung miteinander verbinden. Die resultierende Grafik ähnelt einer topografischen Karte mit Höhenlinien. Die Höhenunterschiede entsprechen den Schönheitsbewertungen.

Die multidimensionale Skalierung wurde für zwei Begriffsgruppen durchgeführt – einerseits häufig vorkommende Begriffe des Natur- und Kulturraums und andererseits häufige Adjektive. Im ersten Fall kamen 56 % der möglichen Begriffskombinationen vor; im zweiten Fall 61.5 %. In den Daten nicht vorkommende Begriffspaare haben in der normalisierten Distanzmatrix die maximale Distanz von 1 zueinander.

5.4 Lexikonerstellung

5.4.1 Unigramme

Die Bilder von ScenicOrNot haben neben einer Landschaftsbewertung mehrheitlich auch einen Begleitkommentar. Aus diesen beiden Informationen lässt sich die durchschnittliche Bewertung jedes verwendeten Wortes ermitteln. Dies kann man berechnen, indem man

für jedes vorkommende Wort dessen jeweilige Bewertungen addiert und zum Schluss den Durchschnitt kalkuliert. Um die Vergleichbarkeit der Worte und die Trefferquote zu erhöhen, wurden alle Worte mit dem oben beschriebenen Lemmatisierer in ihre Grundform gebracht. Zusätzlich wurde bei diesen Berechnungen eine Toponymerkennung (siehe Abschnitt 5.2.1) angewendet, die entsprechende Begriffe als Toponym markiert. Schliesslich werden häufige englische Wörter (siehe Abschnitt 9.1 auf Seite 108) mit Hilfe einer Look-up-Tabelle markiert, damit diese bei den Analysen ausgeschlossen werden können, weil sie bezüglich der Landschaftsbewertung keinen Informationsgehalt haben.

Diese Berechnungen wurden mehrmals durchgeführt. Einerseits wurden die Durchschnittswerte für alle Wörter mit Berücksichtigung aller Kommentare berechnet. Diese Werte wurden für die Erstellung der Wordles, die in Abschnitt 6.1 auf Seite 63 beschrieben sind, verwendet. Dies erlaubt einen Überblick über die Kommentare. Zusätzlich wurden die bewerteten Bilder zufällig in zwei Gruppen aufgeteilt und für beide Gruppen jeweils separat dieselben Berechnungen durchgeführt. Dies dient zum einen dazu, dass die Hälfte der Bilder (also jeweils 105'690) zur Berechnung eines Lexikons für landschaftsspezifische Schönheitswerte benutzt werden kann, das dann auf die andere Hälfte angewendet werden und so validiert werden kann. Zum anderen kann durch die Berechnung zweier Lexika die Stabilität der Lexikonwerte untersucht werden, weil für eine Anzahl häufig verwendeter Wörter, die in beiden Lexika vorkommen, zwei verschiedene Durchschnittswerte vorliegen, die mit einander verglichen werden können.

Nutzergetrennte Lexika Beim zufälligen Aufteilungsprozess werden die Bilder aller Nutzer, die mehr als ein Bild beigesteuert haben, mit grosser Wahrscheinlichkeit in beide Samples verteilt. Dies kann einen Einfluss auf das Verhältnis zwischen den beiden generierten Lexika haben, weil, wie in Abschnitt 3.1.1 auf Seite 23 gezeigt wurde, sich viele Fotografen auf ein Gebiet konzentriert haben. Um dies zu untersuchen, werden neben den zufälligen zwei neue Lexika generiert, deren Eigenschaft darin besteht, dass alle Bilder und folglich auch die Kommentare eines Nutzers innerhalb eines Samples zu finden sind. Dazu wird zuerst für jeden Nutzer von Geograph ermittelt, wie viele Bilder er beigetragen hat. Anschliessend werden die Nutzer zufällig gemischt und dann solange durch die zufällig angeordneten Nutzer iteriert und dabei die jeweiligen Bilder dem einen Sample hinzugefügt, bis ein Grenzwert – die Hälfte aller Bilder – erreicht ist. Mit diesem Vorgehen wird jedes Bild der einen oder anderen Gruppe zugeteilt.

Übereinstimmung der zufälligen Unigramm-Lexika

Die Gegenüberstellung der durchschnittlichen Werte der beiden Lexika ist in den Abbildungen 5.2, 5.3 und 5.4 zu sehen. Der Unterschied dieser Grafiken besteht darin, dass

die Häufigkeit, wie oft ein Wort vorkommen muss, zunimmt. Nicht berücksichtigt werden Toponyme und sehr geläufige Wörter wie *the*. In Abb. 5.2 muss ein Wort in beiden Lexika mindestens 2 Mal in den Kommentaren vorkommen, was auf 10'732 verschiedene Begriffe zutrifft. Der Korrelationskoeffizient r^2 beträgt 0.344. Wie die Tab. 5.1 zeigt, nimmt mit zunehmendem Mindestvorkommen der Wörter die Anzahl der entsprechenden Wörter ab, aber gleichzeitig nimmt r^2 zu. Das bedeutet, dass bei häufiger vorkommenden Wörtern die Übereinstimmung zwischen den beiden Lexika besser wird, was den Erwartungen entspricht. Wie bereits erwähnt sind hier 120 sehr häufige Wörter (siehe Abschnitt 9.1) nicht berücksichtigt. Ausserdem zeigt dieser Zusammenhang, dass für weitergehende Betrachtungen der Wortbewertungen nur solche mit einem gewissen Mindestvorkommen verwendet werden sollten, damit die Resultate stabil sind. Es wird auch deutlich, dass bei zunehmendem Mindestvorkommen die Zahl der entsprechenden Wörter rasch abnimmt, da es sich um eine Zipfverteilung handelt.

Die Stabilität der beiden Lexika zeigt sich auch dadurch, dass sich die Gesamtanzahl der gefundenen Wörter nur um 358 Wörter unterscheidet und auch die Anzahl der gefundenen Toponyme beträgt in beiden Fällen jeweils rund 30'000, obwohl die beiden Stichproben völlig zufällig generiert wurden. Knapp die Hälfte der gefundenen Begriffe sind in den beiden Lexika Toponyme.

Mindestvorkommen	Anzahl Begriffe	r^2
2	10'732	0.344
10	4'549	0.6609
25	2'705	0.8044

Tabelle 5.1: Bestimmtheitsmass zwischen dem Trainings- und dem Validierungsset mit zunehmendem Mindestvorkommen der Begriffe.

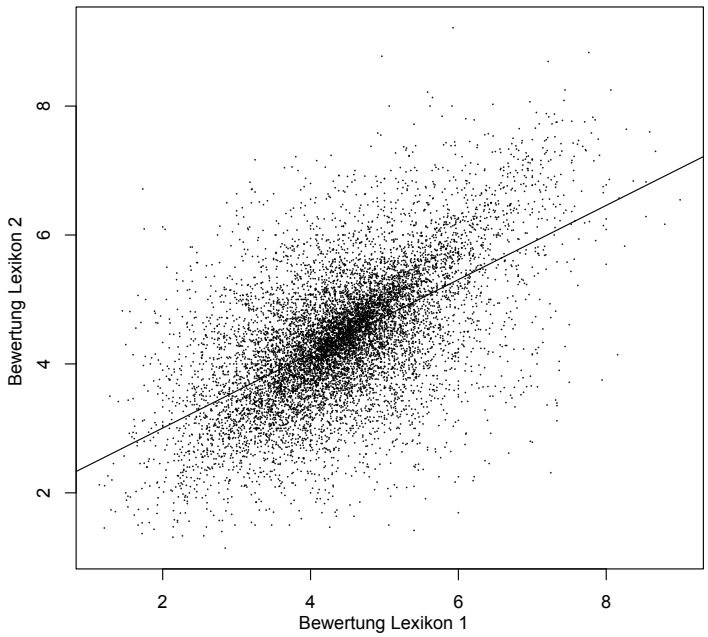


Abb. 5.2: Scatterplot der zwei zufälligen Lexika. Das Bestimmtheitsmass r^2 beträgt 0.344, berücksichtigt wurden alle Wörter, die mindenstens 2 mal in jedem Lexikon vorkommen, was 10'732 Wörtern entspricht.

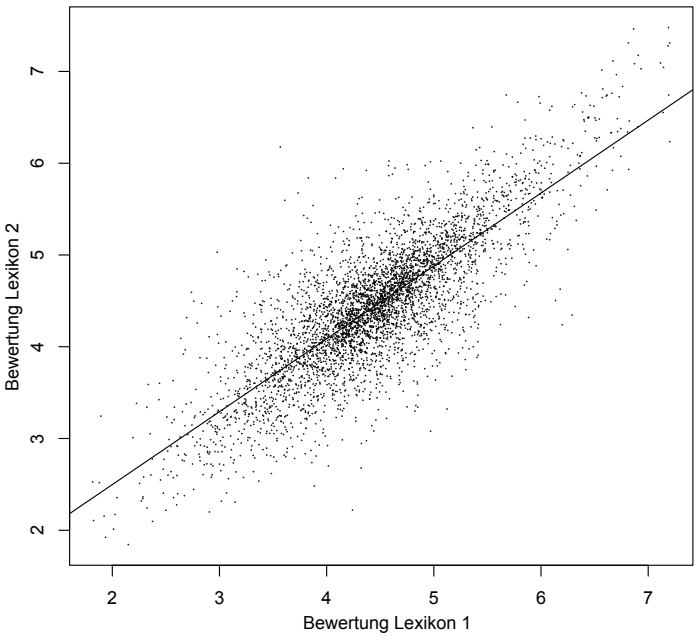


Abb. 5.3: Scatterplot der zwei zufälligen Lexika. Das Bestimmtheitsmass r^2 beträgt 0.6609, berücksichtigt wurden alle Wörter, die mindenstens 10 mal in jedem Lexikon vorkommen, was 4'549 Wörtern entspricht.

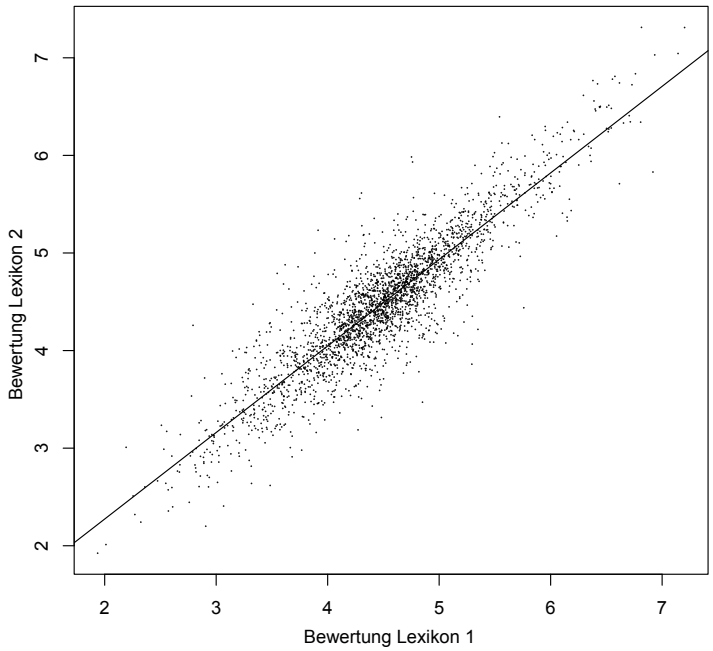


Abb. 5.4: Scatterplot der zwei zufälligen Lexika. Das Bestimmtheitsmass r^2 beträgt 0.8044, berücksichtigt wurden alle Wörter, die mindenstens 25 mal in jedem Lexikon vorkommen, was 2'705 Wörtern entspricht.

Übereinstimmung der nach Nutzer getrennten Lexika

Die Tab. 5.2 auf der nächsten Seite zeigt die Bestimmtheitsmasse zwischen den beiden Lexika, bei denen die Geograph-Nutzer jeweils nur einem Lexikon zugeteilt wurden. Die Werte für r^2 sind bei allen drei untersuchten Mindestvorkommen leicht unter den Werten der zufälligen Lexika (siehe Tab. 5.1). Daraus kann man schliessen, dass ein Teil der Korrelation der zufälligen Lexika dadurch zustande kommt, dass sich Begleittexte von sehr aktiven Nutzern über beide Samples verteilen und dann offenbar ähnlich bewertet werden. Mit der Einschränkung, dass sämtliche Kommentare eines Nutzers in nur einem Sample sind, kann ein Einfluss dieses Effekts ausgeschlossen werden. Ausserdem hat diese Einschränkung zur Folge, dass die Anzahl Begriffe, die in nur einem Lexikon vorkommen, grösser als bei den zufällig aufgeteilten ist. Vor allem für Begriffe, die nur sehr selten verwendet werden, steigt die Wahrscheinlichkeit, dass sie in nur einem Lexikon erscheinen. Insgesamt zeigt die Tab. 5.2 eine etwas schlechtere Übereinstimmung der durchschnittlichen Bewertungen in den beiden nutzergetrennten Lexika, aber ab einem Mindestvorkommen von 25 immer noch ziemlich stabile Berechnungen.

Mindestvorkommen	Anzahl Begriffe	r^2
2	10'121	0.286
10	4'338	0.6171
25	2'572	0.7501

Tabelle 5.2: Bestimmtheitsmasse zwischen den zwei nutzergetrennten Unigramm-Lexika mit zunehmendem Mindestvorkommen der Begriffe

Verteilung der Bewertungen in den zufälligen Unigramm-Lexika

Abb. 5.5 auf der nächsten Seite zeigt eine Gegenüberstellung der Verteilungen der Bewertungen der beiden zufällig generierten Lexika, wiederum ohne Berücksichtigung von Toponymen. Einerseits fällt die gute Übereinstimmung der Verteilung in den beiden Lexika auf, was wiederum ein Zeichen für die Stabilität ist. Andererseits widerspiegeln die beiden Kurven die leicht rechtsschiefe Verteilung der Landschaftsbewertungen, die in Abb. 3.6 auf Seite 29 dargestellt ist. Die rechtsschiefe Verteilung mit dem flachen rechten Ende zeigt, dass Begriffe mit tiefen Bewertungen wesentlich häufiger vorkommen als sehr hohe Bewertungen, weil es weniger Bilder in den hohen Kategorien gibt. In Übereinstimmung mit den Scatterplots weiter oben zeigt die Abb. 5.5, dass die Mehrzahl der Durchschnittsbewertungen der Begriffe zwischen 3 und 6 liegen. Die mit der Anzahl der Begriffe gewichtete Gesamt-Durchschnittsbewertung beträgt im Lexikon 1 4,53 und im Lexikon 2 4,54.

Die Abb. 5.5 berücksichtigt die unterschiedliche Anzahl Bilder pro Kategorie nicht und widerspiegelt darum die Verteilung der Bilder. Die Abb. 5.6 hingegen zeigt die Anzahl Begriffe jeder Kategorie im Verhältnis zur Anzahl Bilder der entsprechenden Kategorie, jeweils unter Ausschluss der Toponyme. Auffällig ist hier, dass trotz der Normalisierung der Anteil negativer Begriffe höher ist als der Anteil positiv bewerteter Begriffe. Die Abb. 5.6 steht deshalb im Widerspruch zur Pollyanna-Hypothese, die besagt, dass die menschliche Sprache eine Verzerrung ins Positive habe. Man würde aufgrund dieser Hypothese eine linksschiefe Verteilung erwarten.

Wenn man aber die in den Kommentaren vorkommenden Begriffe mit ihrer Häufigkeit gewichtet und die entsprechenden Stimmungswerte von SentiWordNet der Begriffe aufsummiert, erhält man eine positive Zahl. Dies bedeutet, dass die Kommentare von Geograph dem oft beobachteten Muster der positiven Verzerrung in der menschlichen Äusserung folgen, wenn die Stimmung der einzelnen Wörter mit einem verbreiteten, allgemeinen Stimmungslexikon berechnet wird.

Es gibt aber eine Diskrepanz zwischen den Schönheitswerten, die die Begriffe durch die Auswertung der Landschaftsbewertungen erhalten und den Werten, die in SentiWordNet

vermerkt sind. Es kann also mit einer positiven Sprache über eine ästhetisch wenig gefällige Landschaft gesprochen werden. Dies kann als Hinweis darauf gedeutet werden, dass zur Beantwortung der zweiten Forschungsfrage, nämlich der Landschaftsbewertungsschätzung, ein allgemeines Lexikon wie SentiWordNet ungeeignet ist.

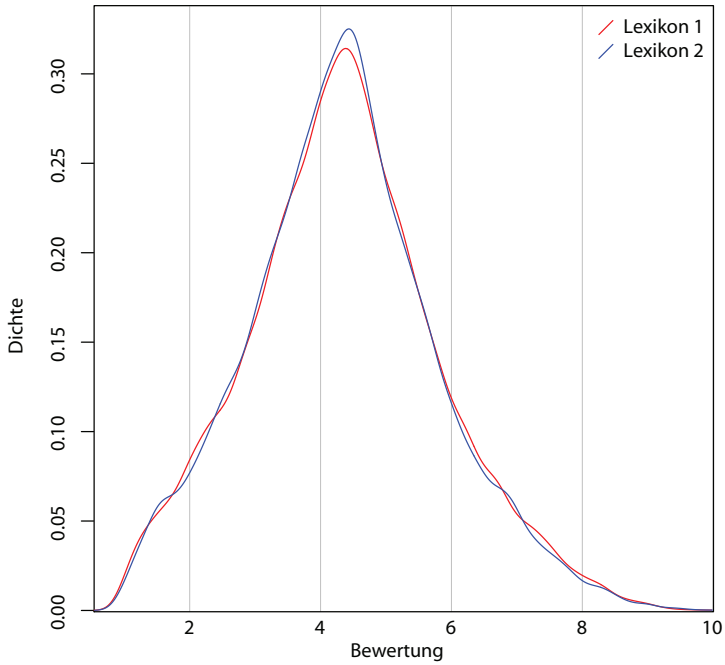


Abb. 5.5: Gegenüberstellung der Dichteverteilung der Bewertungen der Begriffe in den beiden Lexika.

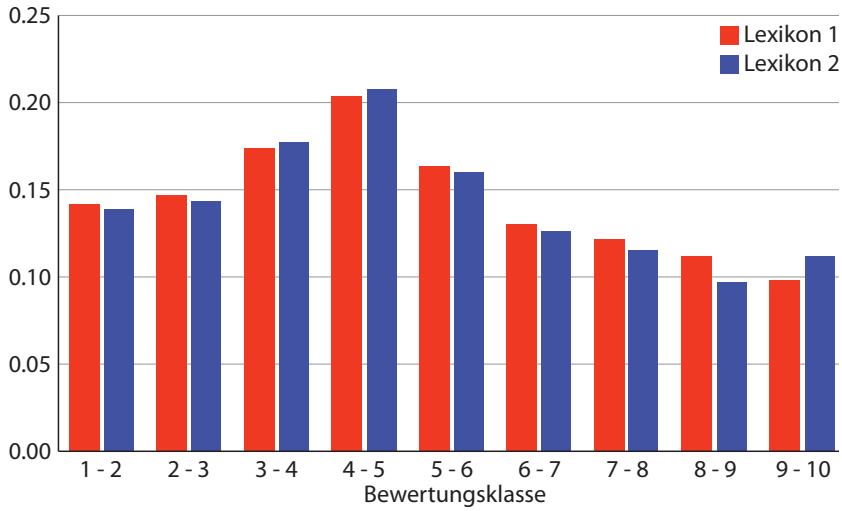


Abb. 5.6: Verhältnis der Anzahl Begriffe und Anzahl Bilder pro Kategorie pro Bewertungskategorie.

Bewertung häufiger Begriffe

Purves et al. (2011) haben die Verwendung von Begriffen in Geograph analysiert und unter anderem eine Liste häufig verwendeter Begriffe generiert, welche in die Kategorien «Elemente», «Qualitäten» und «Aktivitäten» aufgeteilt wurden. Diese sind in Tab. 5.3 wiedergegeben und um ihre durchschnittliche Bewertung aufgrund der oben beschriebenen Methode ergänzt - jeweils für beide Lexika. Die meisten dieser Bewertungen bewegen sich in der Nähe des arithmetischen Mittels, was, unter Berücksichtigung von Abb. 5.7 auf der nächsten Seite, wenig überraschend ist. Obwohl die Bandbreite der Werte in der Tab. 5.3 relativ gering ist, lässt sich ein Muster erkennen: Begriffe, die mit der Natur assoziiert werden (z. B. *hill*, *river*), erzielen durchgehend höhere Bewertungen als Ausdrücke des urbanen Raums (z. B. *road*, *park*). Dieses Muster ist vor allen Dingen bei den «Aktivitäten» und den «Elementen» ausgeprägt, in geringerem Mass aber auch bei den «Qualitäten».

Elemente			**Qualitäten**			**Aktivitäten**		
Begriff	Lex. 1	Lex. 2	Begriff	Lex. 1	Lex. 2	Begriff	Lex. 1	Lex. 2
road	3.99	4.01	old	4.34	4.36	walk	5.05	5.11
farm	4.27	4.27	new	3.81	3.80	grazing	5.21	5.26
lane	4.12	4.13	built[a]	3.91	3.98	running	4.78	4.83
church	4.21	4.18	centre	4.18	4.22	golf	4.17	4.21
bridge	4.37	4.39	square	4.64	4.69	work	3.72	3.84
hill	5.27	5.24	small	4.90	4.82	cycle	4.18	4.10
river	5.27	5.06	water	4.92	4.93	fishing	5.38	5.14
house	4.10	4.05	wood	4.10	4.05	construction	3.82	3.80
park	3.94	3.93	high	4.94	4.92	run	4.62	4.61
street	3.24	3.28	main	3.92	3.87	walking	5.38	5.37

[a] Wegen der Verwendung eines Lemmatizers wird hier der Wert für «build» angegeben.

Tabelle 5.3: Durchschnittliche Bewertung der häufigsten Begriffe in Geograph

Abb. 5.7 auf der nächsten Seite zeigt den Vergleich der Dichteverteilung der Bewertung der häufigsten Begriffe aufgrund der Zusammenstellung von *Purves et al.* (2011). Grundlage sind jeweils gut 500 Begriffe und nicht alle 560 Wörter der ursprünglichen Liste, weil diese auch flexierte Wortformen enthält, welche in den beiden Lexika wegen der Verwendung des Lemmatisierers nicht vorkommen. Im Vergleich zur analogen Darstellung aller Begriffe (Abb. 5.5) zeigt die Betrachtung häufig auf Geograph verwendeter Begriffe eine nahezu symmetrische, annähernd normalverteilte Dichteverteilunsgkurve für beide Lexika. Wiederum sind extreme Bewertungen recht selten und die Spannweite ist geringer als bei Berechnung mit allen Begriffen: Der niedrigste Wert beträgt 1.83, der höchste 7.40.

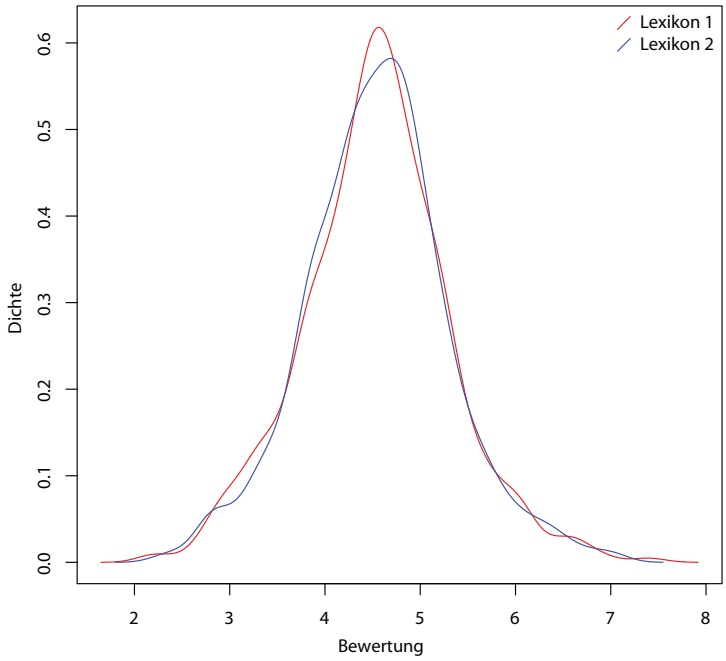


Abb. 5.7: Gegenüberstellung der Dichteverteilung häufiger Begriffe in den beiden Lexika.

Vergleich mit SentiWordNet-Werten

In Tab. 5.3 auf der vorherigen Seite werden die durchschnittlichen Bewertungen der häufigsten Begriffe aus den Unigramm-Lexika gezeigt. Diese können mit den in SentiWordNet verfügbaren Werten verglichen werden. Sämtliche zehn Begriffe, die den «Elementen» zugeordnet wurden, haben in SentiWordNet einen neutralen Stimmungswert, da diese Nomen in der Sprache offenbar keine Stimmung transportieren und von SentiWordNet als vollkommen objektiv eingestuft werden. Bereits im Abschnitt 2.3.4 auf Seite 18 wurde erwähnt, dass nur circa 10 % der Nomen in SentiWordNet nicht objektive Werte aufweisen. Wenn man aber über eine Landschaft spricht, lässt sich aus der Verwendung solcher Begriffe auf die wahrgenommene Schönheit der Landschaft schliessen, weil beispielsweise «road» in Beschreibungen von ästhetisch weniger gefälligen Orten verwendet wird als «river». SentiWordNet hingegen wurde mit der Intention, die Stimmung des Schreibenden zu ergründen, entwickelt.

Anders sieht es bei den beiden Kategorien «Qualitäten» und «Aktivitäten» aus. Die meisten dieser Begriffe haben auch in SentiWordNet keine neutralen Stimmunsgwerte, aber es lässt sich keine Korrelation zwischen der durchschnittlichen Schönheitsbewertung bezüglich der Landschaft und den SentiWordNet-Werten feststellen.

In der Abb. 5.8 auf der nächsten Seite ist eine Gegenüberstellung der Stimmungswerte aus SentiWordNet und der durchschnittlichen Bewertung (gemäss Berechnung in Abschnitt

5.4) einzelner Worte des zufälligen Lexikons dargestellt. Die Grafik führt deutlich vor Augen, dass es zwischen den beiden Datensätzen keine Korrelation gibt. 61.7 % der mit dem Lexikon übereinstimmenden Begriffe haben in SentiWordNet einen neutralen Stimmungswert und werden folglich als objektiv eingestuft.

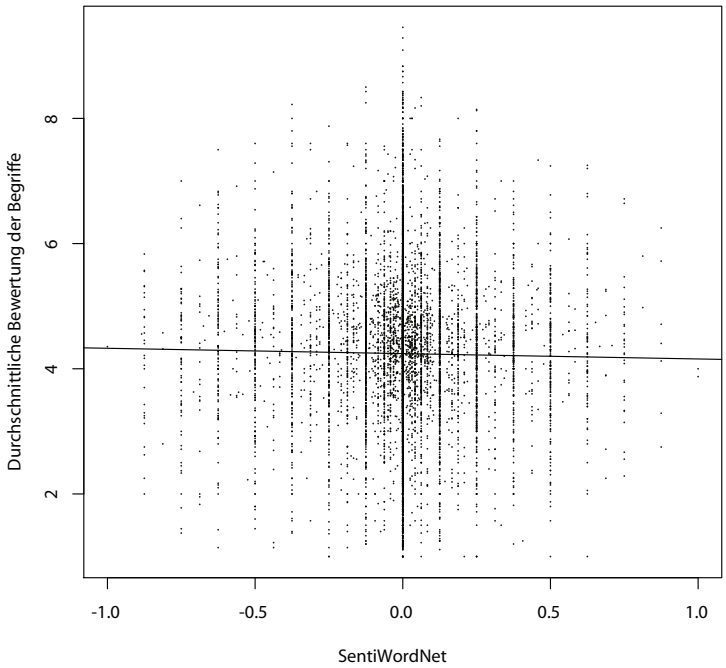


Abb. 5.8: Scatterplot der SentiWordNet-Werte und der Durchschnittswerte im zufälligen Lexikon. Der Korrelationskoeffzient r^2 beträgt 0.0001

All dies legt nahe, dass zur Schätzung einer Landschaftsbewertung auf der Basis eines Textes SentiWordNet als Grundlage ungeeignet ist. Neben den beschriebenen Unigramm-Lexika werden deshalb weitere Lexika erstellt, auf deren Basis eine Schätzung der Landschaftsbewertung vorgenommen werden soll. Zur Lexikonbildung wird jeweils nur die Hälfte der Geograph-Bilder einbezogen, damit eine Validierung vorgenommen werden kann. Diese Lexika werden im Folgenden beschrieben.

5.4.2 Lokales Unigramm-Lexikon

Im Abschnitt 3.1.1 auf Seite 23 wurde dargelegt, dass die Verteilung der 35 aktivsten Geograph-Fotografen räumlich geclustert ist (vgl. Abb. 3.2). Die Werte für die oben beschriebenen Unigramme können deshalb nicht nur global (also für ganz Grossbritannien) berechnet werden, sondern bei Berücksichtigung der einzelnen Fotografen auch lokal. Jedes Unigramm, das in verschiedenen Regionen verwendet wurde, hat dann mehrere Werte, die mit dem Schwerpunkt des betreffenden Geograph-Fotografen verknüpft werden können. Da nicht alle Begriffe von allen Fotografen verwendet wurden, haben nicht alle

Begriffe 35 Werte. Es gibt Begriffe, die nur von einem Fotografen verwendet wurden. Des Weiteren ist die gesamte Anzahl der erfassten Begriffe kleiner als im oben beschriebenen Unigramm-Lexikon, weil die Anzahl analysierter Fotografien mit 65'654 kleiner als bei den zufälligen Unigramm-Lexika ist.

5.4.3 Bigramme

Bei der Erstellung des Unigramm-Lexikons wurde jedes Wort einzeln analysiert. Die nächste Stufe ist die Bildung von Bigrammen. Diese bestehen aus Wortpärchen, die im Text jeweils aneinander grenzen. Wie bei den Unigrammen wurde durch die zufällige Auswahl der Hälfte der ScenicOrNot-Bilder ein Trainingsset geschaffen, damit das Lexikon anhand der anderen Hälfte validiert werden kann. Auch bezüglich der Ermittlung der Durchschnittswerte ist das Prozedere analog zur oben geschilderten Berechnungsweise. Die Bigramme bestehen jeweils aus den Grundformen der gefundenen Worte, weil diese mit dem Lemmatisierer umgeformt worden sind. Auch die Toponyme wurden mit derselben Methode wie im Unigramm-Lexikon gesucht und entsprechend markiert, damit sie ausgefiltert werden können. Die Toponyme können naturgemäss auch aus nur einem Wort bestehen. Die Motivation für dieses Vorgehen ist, dass zusammengehörende Wortpaare dadurch treffendere Bewertungen erhalten. Das folgende Beispiel soll das illustrieren: *nice* hat im Unigramm-Lexikon eine Bewertung von 4.65, *rural* erzielt 3.91 – im gewichteten Durchschnitt ergibt dies einen Wert von 4.27. Im Bigramm-Lexikon dagegen erreicht die Kombination *nice rural* einen Wert von 5.875, was intuitiv sinnvoller erscheint. Durch die Kombinationsmöglichkeiten, die sich bei den Bigrammen ergeben, ist dieses Lexikon wesentlich grösser als das Unigramm-Lexikon, zugleich haben die einzelnen Einträge aber oft geringere Häufigkeiten. Insgesamt beinhaltet das Bigramm-Lexikon Durchschnittswerte für gut 300'000 Bigramme. Wegen des teilweise seltenen Vorkommens einzelner Bigramme hängen die Bewertungen stark von wenigen Bildern ab.

5.4.4 Trigramme

Das Trigramm-Lexikon baut auf dreiteiligen Wortgruppen auf, wobei die Worte wiederum aufeinander folgen müssen. Es stellt damit die nächste Stufe nach dem Bigramm-Lexikon dar. Auch hier wurden die durchschnittlichen Bewertungen aller Wortgruppen berechnet. Das Trainingsset ist dasselbe, das auch für die Uni- und die Bigramme verwendet wurde. Die Toponyme, die auch hier aus nur einem Wort bestehen können, wurden auch im Trigramm-Lexikon detektiert. Die einzelnen Worte der Trigramme wurden mit dem Lemmatisierer in ihre Grundform gebracht, damit bei der Schätzung später mehr Treffer möglich sind. Die Aufteilung der Kommentare in Dreiergruppen erweitert die Kombinati-

onsmöglichkeiten und die Anzahl der Einträge steigt auf über 700'000 Wortgruppen. Wie auch beim Bigramm-Lexikon kommen viele Kombinationen im gesamten Trainingsset nur ein einziges Mal vor.

5.4.5 Kookkurrenz

Kookkurrenz (engl. *co-occurrence*) bezeichnet das gleichzeitige Vorkommen bestimmter Begriffe innerhalb eines Korpus, wobei im Gegensatz zu den Bigrammen die Reihenfolge egal ist und die Begriffe auch nicht nebeneinander stehen müssen. Der Korpus kann entweder aus einem ganzen Dokument bestehen oder auch einzelne Sätze des Dokumentes beinhalten. Nach *Matsuo und Ishizuka* (2004) wird hier bei der Lexikongenerierung die Kookkurrenz auf Satzebene verwendet. Weiter wird nicht die Kookkurrenz aller Begriffe betrachtet, sondern die im Abschnitt 5.4.1 auf Seite 49 beschriebenen häufigen Begriffe von *Purves et al.* (2011). Es wurde das gleichzeitige Auftreten der «Elemente» mit den «Qualitäten» untersucht. Bei diesen häufigen Begriffen sind 348 der Kategorie «Elemente» und 266 der Kategorie «Qualitäten» zugeordnet. Insgesamt gibt es folglich 78'648 Kombinationen, wie diese Begriffe zusammen in einem Satz auftreten können, wenn die Reihenfolge nicht berücksichtigt wird. Allerdings kommen im verwendeten Datensatz, der die Hälfte der von ScenicOrNot bewerteten Fotos enthält (zufällige Auswahl), nicht alle diese Kombinationen vor, sondern nur 42'088. Ferner sind 37 Begriffe in beiden Kategorien vorhanden (bspw. *skyline*), was bedeutet, dass sie mit sich selbst kombiniert werden könnten. Weil dies aber nicht erwünscht ist, bleiben schlussendlich 42'052 Begriffskombinationen, für die jeweils alle vergebenen Stimmen zusammengezählt wurden. Damit kann die durchschnittliche Bewertung jeder Kombination berechnet werden. Von allen Kombinationen, die mehr als zehn Mal gefunden wurden, erzielt «common» «motorway» (1.36) die schlechteste Bewertung – «sandy» «loch» (7.58) hingegen die beste.

5.4.6 Varianz in den Lexika

In jedem der oben beschriebenen Lexika kann aus den einzelnen abgegebenen Stimmen die Varianz für jeden individuellen Eintrag gemäss der Formel 5.1 berechnet werden.

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{5.1}$$

Begriffe, bei deren Bewertung grosse Einigkeit herrscht, haben eine kleine Varianz. Dies kann bei der Schätzung einer Landschaftsbewertung zur Gewichtung genutzt werden, indem Begriffe mit kleiner Varianz ein grösseres Gewicht erhalten als solche mit grosser Varianz, weil die Sicherheit der Bewertung grösser ist. Eine geringe Anzahl Begriffe, die

sehr selten vorkommen, haben eine Varianz von 0, weil sie nur Stimmen einer einzigen Bewertungskategorie erhalten haben. Die Mehrheit der Begriffe hat aber eine Varianz der Bewertungen, die grösser als 1 ist.

5.5 Schätzung einer Landschaftsbewertung

5.5.1 Vorbemerkungen

Mit den obigen Ausführungen wurde dargelegt, dass sich die Verwendung des Stimmungslexikons SentiWordNet kaum für die Bewertung von Texten über Landschaften eignet, weil nicht die Stimmung des Schreibenden ergründet werden soll, sondern die Qualität der beschriebenen Landschaft. Die Landschaftsbeschreibungen bestehen oft aus Beschreibungen der vorkommenden Elemente, die Gefühle der Schreibenden hingegen sind kaum vorhanden, wie sich beispielsweise bei der Analyse der meistverwendeten Begriffe zeigt. Die anfängliche Idee, zur Schätzung der Landschaftsqualität SentiWordNet zu verwenden, kann deshalb nicht verfolgt werden. Stattdessen bietet es sich an, ein auf anderen Landschaftsbewertungen aufbauendes Lexikon zu verwenden. Im Abschnitt 2.3 wurde erwähnt, dass für viele Opinion-Mining-Aufgaben die Bestimmung der Subjektivität von Texten wichtig ist. Die Ausführungen im Abschnitt 5.4.1 und die Abb. 5.8 zeigen aber, dass die Mehrzahl der Begriffe, die in den Geographkommentaren vorkommen, objektiv sind. Bei den weiteren Rechnungen wird deshalb nicht berücksichtigt, inwiefern die Texte subjektiv oder objektiv sind.

Im oben beschriebenen Vorgehen wurden verschiedene umfassende Lexika mit Schönheitswerten für einzelne Begriffe erstellt. Weil nur die Hälfte der Bilder zur Erstellung eines Lexikons verwendet wurden, kann mit der verbleibenden Hälfte die Schätzung der Landschaftsästhetik verifiziert werden.

Die eine Hälfte der Daten wird also zum Training des Bewertungslexikons verwendet – die andere Hälfte, für die die wahren Bewertungen ebenfalls vorliegen, wird zur Validierung gebraucht, indem der geschätzte mit dem wirklichen Wert verglichen wird.

Es ergeben sich unterschiedliche Methoden, wie die Landschaftsbewertungen geschätzt werden können:

- Schätzung mit Unigrammen
- Schätzung mit lokalen Unigrammen
- Schätzung nur mit häufigen Begriffen
- Schätzung mit Kookkurrenz-Lexikon
- Schätzung mit Bigrammen
- Schätzung mit Trigrammen

Diese verschiedenen Schätzmethoden können ausserdem mit einer Toponymfilterung und einer Synonym-Suche (ausser bei Bi- und Trigrammen) kombiniert werden. Zudem besteht die Möglichkeit, die Schönheitswerte vor der Schätzung zu transformieren. Die Abb. 5.9 zeigt schematisch das Vorgehen zur Schätzung der Landschaftsbewertung schematisch.

Um die Schätzung durchzuführen, wählt ein PHP-Skript die relevanten Bilder, die den Kriterien entsprechen aus der Datenbank und spaltet deren Texte in Uni-, Bi- oder Trigramme auf. Anschliessend können die Werte automatisiert im entsprechenden Lexikon nachgeschlagen werden und so die geschätzte Landschaftsbewertung berechnet werden und in der Datenbank abgelegt werden. Ein weiteres Skript kann aus diesen Schätzungen durch den Vergleich mit den wahren Werten den erklärten Anteil berechnen.

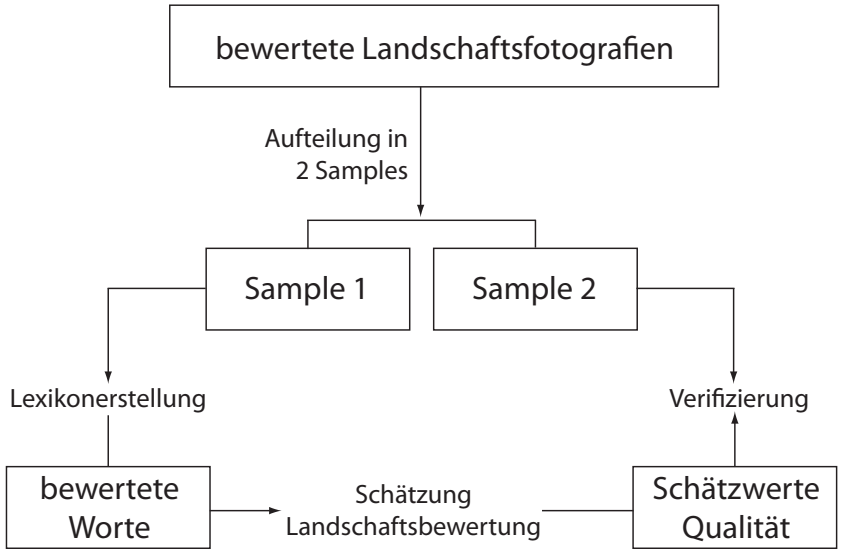
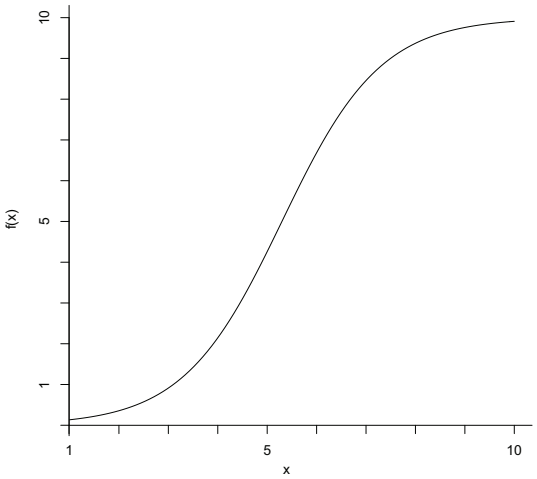


Abb. 5.9: Schematische Darstellung der Vorgehensweise zur Entwicklung des Schätzalgorithmus.

5.5.2 Transformierung der Schönheitswerte

Um die einzelnen durchschnittlichen Schönheitswerte der Uni-, Bi- und Trigramme zu verstärken, kann eine logarithmische Wachstumsfunktion (siehe Gleichung 5.2 und Abb. 5.10) auf den ursprünglichen Wert angewendet werden. Diese Funktion senkt tiefe Bewertungen und erhöht hohe zusätzlich, wobei das Maximum nach wie vor bei 10 liegt. Die Funktion streckt die Werte insbesondere im mittleren Bereich und erhöht so, bildlich gesprochen, den Kontrast. Weil die ursprünglichen Werte in der Datenstruktur erhalten bleiben, ist es möglich, die Wirksamkeit der Transformation zu evaluieren, indem beide Varianten durchgerechnet werden.



$$f(x) = \frac{10}{1 + 200e^{-x}} \tag{5.2}$$

Abb. 5.10: Logarithmische Wachstumsfunktion, die zur Transformation verwendet wurde

5.5.3 Berechnung der Gewichtung

Im Abschnitt 5.4.6 auf Seite 53 wurde die Berechnung der Varianz für die einzelnen Lexikon-Einträge beschrieben. Eine kleine Varianz bedeutet grosse Einigkeit bei der Bewertung des entsprechenden Begriffs und deshalb eine grosse Aussagekraft der Bewertung. Es bietet sich folglich an, die Bewertungen mit der Varianz zu gewichten. Um die Varianz zur Gewichtung zu verwenden, kann aber nicht der Wert direkt verwendet werden, weil eine kleine Varianz ein grosses Gewicht haben soll. Das Gewicht muss deshalb umgekehrt proportional zur Varianz stehen. Dieses Ziel wird mit der Transformierung entsprechend der Formel 5.3 erreicht. Die Berechnung des Gewichts ist eine Funktion der Varianz und der Anzahl der abgegebenen Stimmen. Um bei zunehmender Varianz ein kleineres Gewicht zu erhalten, besteht der Kern der Gewichtungsfunktion aus dem Kehrwert der Varianz. m ist kleiner als 1, um die Abnahme des Gewichts mit zunehmender Varianz zu verlangsamen. Die verwendeten Parameter werden im Abschnitt 5.5.8 erläutert. Die grösste Sicherheit bei der Bewertung eines Begriffs besteht dann, wenn die Varianz auch bei grosser Stimmenzahl klein ist. In der Gleichung wird dem mit dem doppelten natürlichen Logarithmus Rechnung getragen. Dadurch erhöht eine grosse Stimmenzahl das Gewicht. Wenn die Varianz null beträgt, wäre das Gewicht gemäss dieser Formel nicht definiert. Um dies zu verhindern, wird in diesem Fall ein Gewicht von 3 verwendet.

$$\begin{aligned} v = s^2, \quad SdS = Summe\ der\ Stimmen \\ Gewicht(v, SdS) = \frac{1}{v^m} * \ln(\ln(SdS)) \end{aligned} \tag{5.3}$$

5.5.4 Berechnung des Schätzwertes

Der geschätzte Wert einer Landschaftsbewertung berechnet sich als gewichteter Durchschnitt der vorkommenden Wörter im Kommentar gemäss der Formel 5.4. Sehr häufige Wörter wie etwa *the* werden vor der Berechnung herausgefiltert. Je nach Parameterwahl (siehe Abschnitt 5.5.8) nimmt die Anzahl der miteinbezogenen Wörter ab – insbesondere wenn die Anforderung an das Mindestvorkommen der Wörter im Lexikon hoch ist.

$$\text{Schätzwert} = \frac{\sum Bewertung * Gewicht}{\sum Gewicht} \tag{5.4}$$

5.5.5 Schätzung mit den häufigsten Begriffen

Die einfachste Möglichkeit zur Schätzung der Landschaftsbewertung mit einem Lexikon besteht darin, die Kommentare mit den Schönheitswerten der häufigsten Begriffe (gemäss *Purves et al.* (2011)) zu evaluieren. Dabei basieren die Landschaftsschätzungen auf den Durchschnittsbewertungen von lediglich 560 Begriffen.

Die überschaubare Anzahl der Begriffe ermöglicht eine manuelle Bewertung der einzelnen Begriffe, die danach anstelle der aus den Daten generierten verwendet werden kann.

5.5.6 Schätzung mit lokalen Unigrammen

Bei dieser Methode werden die Landschaftsbewertungen mit dem im Abschnitt 5.4.2 beschriebenen räumlich differenzierten Begriffslexikon geschätzt. Die Vermutung ist, dass die Schätzung verbessert werden kann, wenn die Schönheitswerte aus der näheren Umgebung des zu schätzenden Bildes stammen. Dazu wird bei jedem Bild zuerst berechnet, welcher Schwerpunkt der 35 fleissigsten Nutzer der nächste ist (siehe auch Abb. 3.2). Dazu sind als zusätzliche Information die Koordinaten der Bilder nötig. Wenn der nächste Nutzer bestimmt ist, kann im Begriffslexikon der Wert und das Gewicht der Begriffe nachgeschlagen werden – basierend auf den Bildern des entsprechenden Nutzers. Das Vorgehen ist Abb. 5.11 dargestellt.

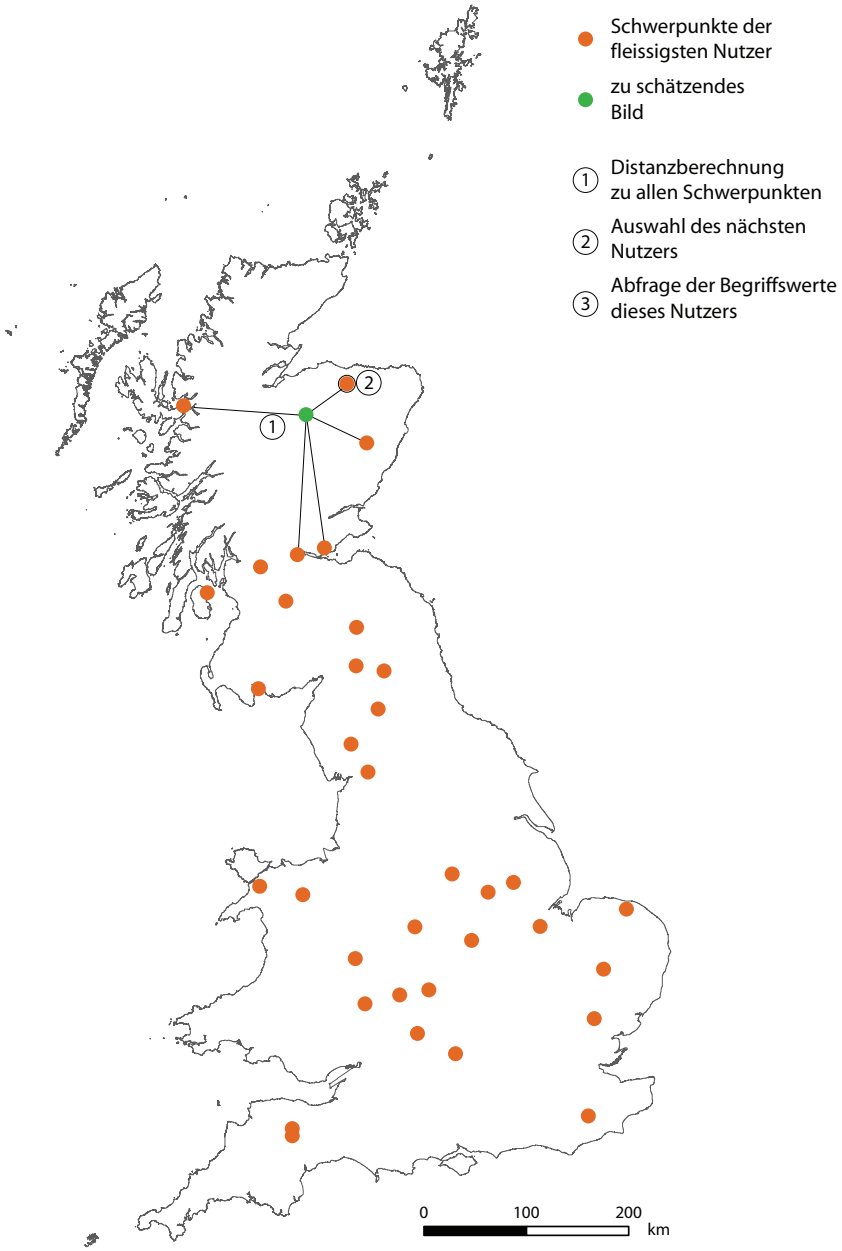


Abb. 5.11: Vorgehen zur lokalen Unigramm-Schätzung mit den Schwerpunkten der fleissigsten Geograph-Fotografen (Quelle Umrisse Grossbritannien: OSGB OpenData)

5.5.7 Schätzung mit Bigrammen und Trigrammen

Wie bereits bei der Lexikonerstellung beschrieben wurde, bestehen Bi- und Trigramme aus Wortgruppen von zwei bzw. drei aufeinanderfolgenden Worten. Um diese Lexika zur Schätzung zu verwenden, werden die Kommentare der zu schätzenden Bilder in Bi- resp. Trigramme aufgespalten und die so entstehenden Wortgruppen werden in der Folge mit dem jeweiligen Lexikon verglichen, um die Werte nachzuschlagen.

5.5.8 Abhängigkeit der Schätzung von der Parameterwahl

Bei der Schätzung der Landschaftsbewertung können neben der Wahl des zugrundeliegenden Lexikons jeweils auch verschiedene Parameter gewählt werden, die einen Einfluss auf das Resultat haben können. Die beeinflussbaren Parameter sind in Tab. 5.4 aufgelistet.

Parameter	**Werte**				
minimale Anzahl Worte im Kommentar	10	20	30	40	50
Mindestvorkommen im Trainingsset	0	5	10	20	25
maximale Varianz der Stimmen	0.75	1.5	2.4	50	
Gewichtungskonstante m (Exponent)	0.1[a]	0.2[a]	0.3[a]	0.4[a]	0.5

[a] nur bei Unigrammen

Tabelle 5.4: Parameter zur Schätzung und ihre Werte

Da *a priori* nicht bekannt ist, welche Parameterkombination die sinnvollste ist und wie sich die Parameterwahl auf die Resultate auswirkt, werden die in Tab. 5.4 aufgelisteten Werte miteinander kombiniert und für jede Kombination jeweils eine Bewertungsschätzung gerechnet. Insgesamt ergeben sich so 500 Kombinationen ($5^3 * 4$) für die Unigramme. Weil sich bei den Unigrammen gezeigt hat, dass der Gewichtungsexponent 0.5 die besten Resultate erzielt, wurde für die Berechnung der Kookkurrenz, Bigramm- und Trigrammwerte nur dieser verwendet. Dadurch ergeben sich hier jeweils 100 Kombinationen ($5^2 * 4$).

Um den Effekt der Transformation der Begriffswerte (siehe Abschnitt 5.5.2) zu analysieren, werden diese Berechnungen einmal mit den ursprünglichen und einmal mit den transformierten Schönheitswerten durchgeführt. Die Resultate dieser Berechnungen werden in der Datenbank abgespeichert, damit die Korrelationsberechnungen später effizient mit einem Skript durchgeführt werden können.

Für jede Kombination werden maximal 2000 zufällig ausgewählte Bilder des Validierungssets verwendet, wobei diese Zahl bei restriktiven Anforderungen an das Validierungsset stark sinkt. So kommen nur 531 Bilder zur Analyse in Frage, wenn der Kommentar mindestens 50 Wörter enthalten soll und die Varianz der Stimmen kleiner als 0.75 sein soll. Die

Parameter *minimale Anzahl Worte im Kommentar* und *maximale Varianz der Stimmen* bedingen jeweils neue Selektionen der zu schätzenden Bilder. Die Durchgänge bei den anderen Parametern werden hingegen mit der jeweils gleichen Bildauswahl durchgeführt. Der Vorteil davon besteht darin, dass so exakt der Effekt des sich ändernden Parameters gemessen werden kann.

Wenn ein Kommentar keine relevanten Begriffe bzw. Kookkurrenzen (nicht im Lexikon erfasst oder nur sehr häufige Begriffe) enthält, wird das entsprechende Bild nicht weiter berücksichtigt. Die Anzahl der einbezogenen Worte sinkt, wenn die Anforderung an das Mindestvorkommen im Trainingsset steigt. Allerdings sind die Werte der oft vorkommenden Begriffe stabiler, wie die Berechnungen im Abschnitt 5.4.1 zeigen. Es wird daher vermutet, dass höhere Mindestvorkommen in den Lexika bessere Resultate erzielen.

5.5.9 Synonymsuche

Nicht alle in einem Beschreibungstext gefundenen Wörter sind notwendigerweise in einem Begriffslexikon vorhanden, insbesondere wenn der Schwellenwert für die Häufigkeit im Trainingsset hoch gelegt wird. Entweder können diese Wörter übergangen werden oder man kann der Problematik begegnen, indem man versucht, ein Synonym zum nicht erfassten Wort zu finden. Die Anforderung an das Synonym ist, dass es im verwendeten Lexikon vorkommt. Folgendes Beispiel soll das verdeutlichen:

Wenn in einem Kommentar das Wort *desolate* vorkommt und dieses aufgrund der gewählten Einschränkungen (Mindestvorkommen) im Lexikon nicht gefunden wurde, kann der Algorithmus nach einem Synonym suchen. Das Problem ist dann, dass oft mehrere Synonyme gefunden werden, da im verwendeten Thesaurus für die meisten Wörter zahlreiche Synonyme vorhanden sind. Die Frage ist deshalb, wie automatisch ein passendes ausgewählt werden kann. In dieser Arbeit werden die gefundenen Synonyme – also Begriffe, die im Lexikon vorkommen – entsprechend ihrem Gewicht sortiert und dasjenige mit dem grössten Gewicht ausgewählt. Beim Wort *desolate* wird mit dieser Vorgehensweise *empty* als Synonym ermittelt.

Im erwähnten Beispiel ist die Auswahl zufriedenstellend, aber aufgrund der Datenmenge ist eine Kontrolle der Synonymauswahl unmöglich und es ist unklar, ob die gewählten Synonyme immer passend sind.

5.5.10 Beispiel einer Bewertungsschätzung

Anhand eines konkreten Beispielbildes (siehe Abb. 5.12) wird nun gezeigt, wie eine Schätzung der Landschaftsbewertung in dieser Arbeit umgesetzt wird.

Abb. 5.12: Beispiel einer Bewertungsschätzung

Auf der linken Seite der Abbildung ist das Bild von Geograph zu sehen, rechts oben der zugehörige Kommentar. Zuerst werden im Text die Toponyme gesucht – im Beispiel wurde nur eines von zwei gefunden. Danach wird jedes Wort lemmatisiert und im Stimmungslexikon gesucht. Wenn es gefunden wird und nicht zu den sehr häufigen Begriffen gehört, wird sein Schönheitswert mit dem Gewicht multipliziert. Nachdem dies für alle Wörter durchgeführt wurde, kann die geschätzte Bewertung durch Aufsummieren und Division durch die Summe der Gewichte berechnet werden.

Im Beispiel ist eine grosse Differenz zwischen der Bewertung von ScenicOrNot und dem Schätzwert zu sehen. Dies bedeutet aber nicht notwendigerweise, dass die Schätzung schlecht ist, denn durch die Gewichtung und vor allem durch die Transformierung der Schönheitswerte ergibt sich eine andere Spannweite der Bewertungen. Die Güte einer Schätzmethode wird deshalb nicht mit einem einzelnen Bild bestimmt, sondern mit einigen hundert bis einigen tausend. Zur Bestimmung der Schätzungsqualität wird der Zusammenhang zwischen der Schätzung und dem wahren Wert mit einem linearen Modell gemessen.

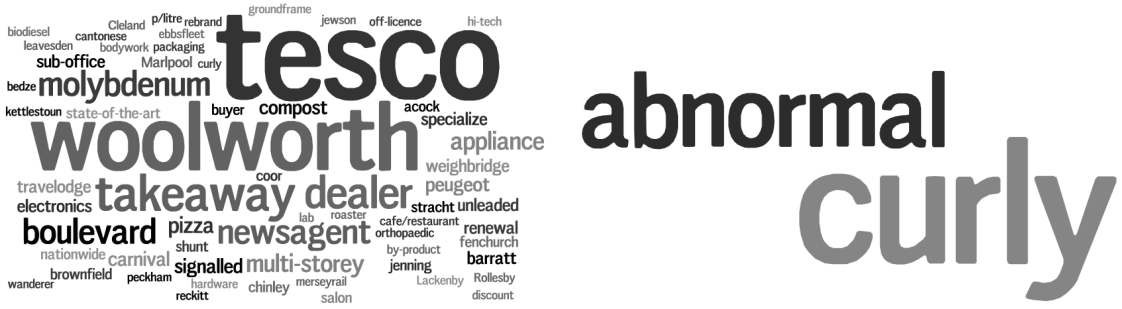
6 Resultate

6.1 Wortwolken

Um einen Überblick über den Inhalt der Kommentare zu erhalten, bietet es sich an Wortwolken, sogenannte Wordles[1], zu erstellen - sowohl für alle Wörter (aber ohne Toponyme und häufige Wörter) als auch nur für Adjektive, die mit einer Look-up-Table markiert wurden. Die Wortwolken sind in 9 Kategorien aufgeteilt (durchschnittliche Bewertung eines Wortes 1 – 2 usw.) und die Schönheitswerte wurden mit den in Abschnitt 5.4 auf Seite 42 beschriebenen Methoden berechnet. Die Wortwolken basieren auf der Auswertung aller Bilder, nicht eines zufälligen Subsamples, damit ein Überblick über die Art der Kommentare auf Geograph gewonnen werden kann. Diese Wortwolken sind in den Abbildungen 6.1 bis 6.8 dargestellt. Die Grösse der einzelnen Begriffe gibt die relative Häufigkeit zu den anderen Begriffen derselben Kategorie an. Die Adjektive werden nur dargestellt, wenn sie mindestens zwei Mal in den Kommentaren vorkommen, für die Darstellung aller Wörter wurden nur solche einbezogen, die mindestens vier Mal vorkommen. Insbesondere in den höheren Bewertungskategorien gibt es nur wenige Begriffe, die diesen Kriterien genügen und in der obersten Kategorie 9 – 10 trifft es auf keine Begriffe mehr zu, weil die Bewertungen rechtsschief verteilt sind, wie die Abb. 5.5 auf Seite 48 zeigt.

In den tiefen Bewertungskategorien dominieren Begriffe, die mit Vororten assoziiert werden können, u. a. Namen von Supermärkten und Begriffe, die im Zusammenhang mit Tankstellen stehen. Auch Begriffe wie *housing*, *office*, *motorway*, *airport* und dergleichen sind in einer tiefen Bewertungskategorie angesiedelt. Auch die Mehrzahl der Adjektive der tiefen Kategorien hat eine intuitiv negative ästhetische Konnotation. Ab einer durchschnittlichen Bewertung von 4 tauchen vermehrt Begriffe des ruralen Raumes auf, bei den Adjektiven ist dies mehrheitlich erst ab einer Bewertung von 5 der Fall. Sehr häufig kommt auch der Begriff «looking» vor, weil die Fotografien in einfachen Kommentaren oft so beschrieben sind. Bei der Beantwortung der Forschungsfrage 1 kann als Erstes festgehalten werden, dass positiv bewertete Landschaften vermehrt Begriffe des natürlichen Raumes in ihren Legenden enthalten. Die positivere Bewertung der natürlichen Landschaften (siehe *Stadler* (2010)) spiegelt sich also auch in den Begleittexten von Geograph.

[1] Die Wordles wurden mit www.wordle.net erstellt.



(a) alle Wörter (b) Adjektive

Abb. 6.1: Wörter mit einer durchschnittlichen Bewertung von 1 – 2

(a) alle Wörter (b) Adjektive

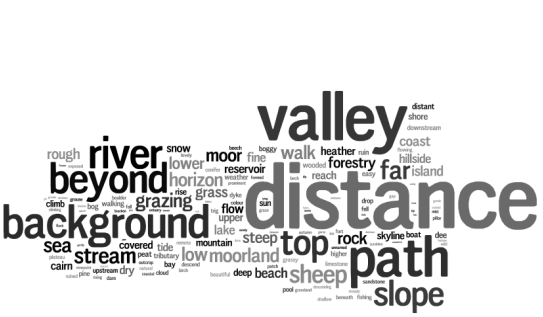
Abb. 6.2: Wörter mit einer durchschnittlichen Bewertung von 2 – 3

(a) alle Wörter (b) Adjektive

Abb. 6.3: Wörter mit einer durchschnittlichen Bewertung von 3 – 4

(a) alle Wörter (b) Adjektive

Abb. 6.4: Wörter mit einer durchschnittlichen Bewertung von 4 – 5

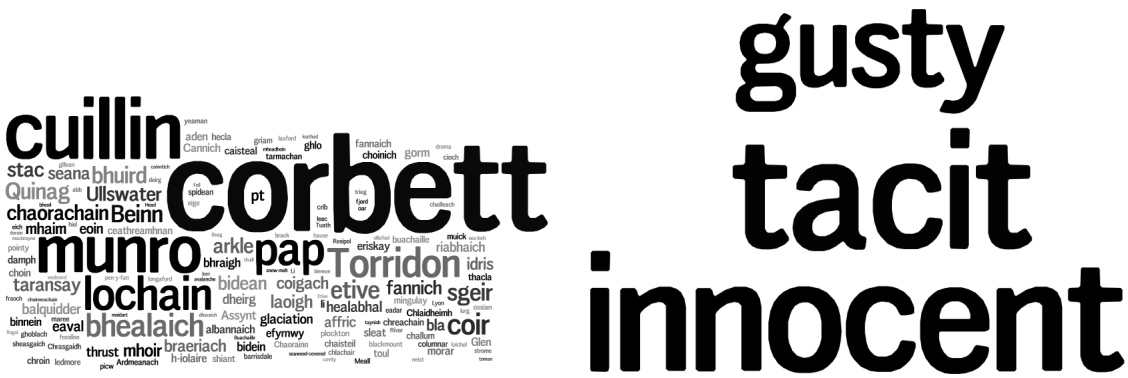


(a) alle Wörter (b) Adjektive

Abb. 6.5: Wörter mit einer durchschnittlichen Bewertung von 5 – 6

(a) alle Wörter (b) Adjektive

Abb. 6.6: Wörter mit einer durchschnittlichen Bewertung von 6 – 7



(a) alle Wörter (b) Adjektive

Abb. 6.7: Wörter mit einer durchschnittlichen Bewertung von 7 – 8

(a) alle Wörter

Abb. 6.8: Wörter mit einer durchschnittlichen Bewertung von 8 – 9 (es gibt keine Adjektive in dieser Kategorie)

Je höher die Bewertungskategorie ist, besonders ab einer Bewertung von 7, desto öfter tauchen naturbezogene Begriffe auf, vor allem zahlreiche spezifisch schottische Ausdrücke für eine Vielzahl von geomorphologischen Formen. Aber auch die Adjektive in den hohen Bewertungskategorien beschreiben in der Mehrheit Dinge im Zusammenhang mit der Natur.

Wie ein genauerer Blick in die Daten demonstriert, werden Wörter teilweise in einem unerwarteten Zusammenhang verwendet. *Stalker* erreicht eine hohe Bewertung (6.73), weil es offensichtlich in seinem ursprünglichen Sinn, nämlich *Pirschjäger*, in ländlichen Gebieten benutzt wird, und nicht etwa negativ konnotiert in eher städtischen Umgebungen, wie man das erwarten könnte. In SentiWordNet hat dieses Wort eine neutrale Bewertung, weil es sowohl negativ als auch positiv verwendet werden kann. Bei diesem Begriff kommt allerdings erschwerend hinzu, dass es in Schottland ein «Castle Stalker» gibt, das im Gazetteer zwar als Toponym vermerkt ist, aber unter Umständen nicht immer als solches erkannt worden ist.

6.1.1 Rangliste höchster und tiefster Bewertungen

Rang	Begriff	Bewertung
1	corbett	7.14
2	coire	6.93
3	stalker	6.73
4	ravine	6.71
5	lochan	6.7
6	clifftop	6.62
7	scree	6.6
8	gneiss	6.56
9	corrie	6.55
10	rocky	6.54
11	basalt	6.52
12	loch	6.52
13	craggy	6.52
14	waterfall	6.503
15	carn	6.501

Rang	Begriff	Bewertung
1	retail	2.01
2	code	2.1
3	recycling	2.19
4	supermarket	2.25
5	shopping	2.27
6	comprehensive	2.28
7	distribution	2.32
8	petrol	2.36
9	warehouse	2.47
10	store	2.5
11	semi-detached	2.55
12	demolition	2.56
13	garage	2.57
14	stadium	2.571
15	boom	2.59

Tabelle 6.1: Rangliste der 15 am besten (links) bzw. schlechtesten (rechts) bewerteten Begriffe

Die Tab. 6.1 zeigt eine Rangliste von 15 sehr gut bzw. sehr schlecht bewerteten Begriffen mit ihren Schönheitswerten. Die Auswahl wurde so getroffen, dass keine Eigennamen wie *Woolworth* (schlecht bewertet) oder *cuillin* (gut bewertet) enthalten sind. Die Begriffe mussten mindestens 10 mal gefunden werden und im Minimum 200 Stimmen erhalten haben.

6.1.2 Multidimensionale Skalierung

Um die in der Tab. 6.1 dargestellten Verhältnisse anschaulich zu visualisieren, wurden zwei multidimensionale Skalierungen erstellt. Die Abb. 6.9 zeigt das Resultat der multidimensionalen Skalierung mit kultur- und naturräumlichen Begriffen. Auf den ersten Blick sind zwei Gruppen erkennbar – auf der linken Seite Begriffe des urbanen, besiedelten Raums und rechts Begriffe mit einem ländlichen Kontext. Weil die Begriffe des urbanen Raums wesentlich schlechtere Bewertungen als die des natürlichen Raums haben, entsteht eine scharfe farbliche Trennung in der Mitte der Karte. Dadurch bildet sich eine ausgeprägte Schelfkante. Auf der rechten Seite – dem Festland – lässt sich ein Verlauf von Begriffen der Küste (oben) zu Begriffen der Bergwelt (unten) feststellen, deren grosse Distanz zueinander dadurch erklärbar ist, dass sie offensichtlich selten gemeinsam in

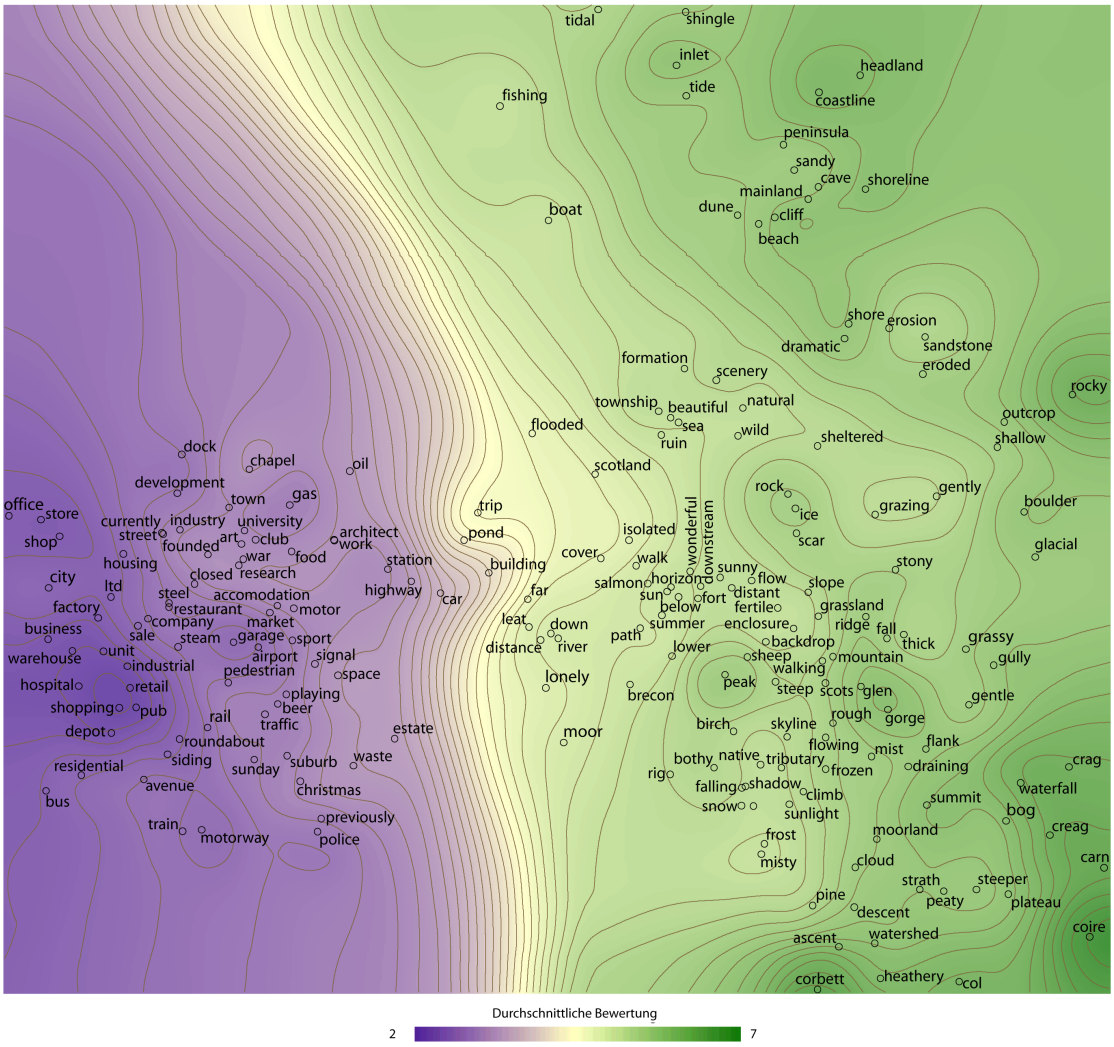


Abb. 6.9: Resultat der multidimensionalen Skalierung. Je näher zwei Begriffe sind, desto häufiger kommen sie zusammen vor. Die Farbe stellt die durchschnittliche Bewertung dar.

Kommentaren auftreten. Innerhalb der urbanen Gruppe besteht die Tendenz, dass oben städtische Einrichtungen und Funktionen zu finden sind, unten hingegen Begriffe des Verkehrs. In der Auswertung der multidimensionalen Skalierung manifestiert sich die bereits in den Wortwolken gefundene Struktur sehr deutlich.

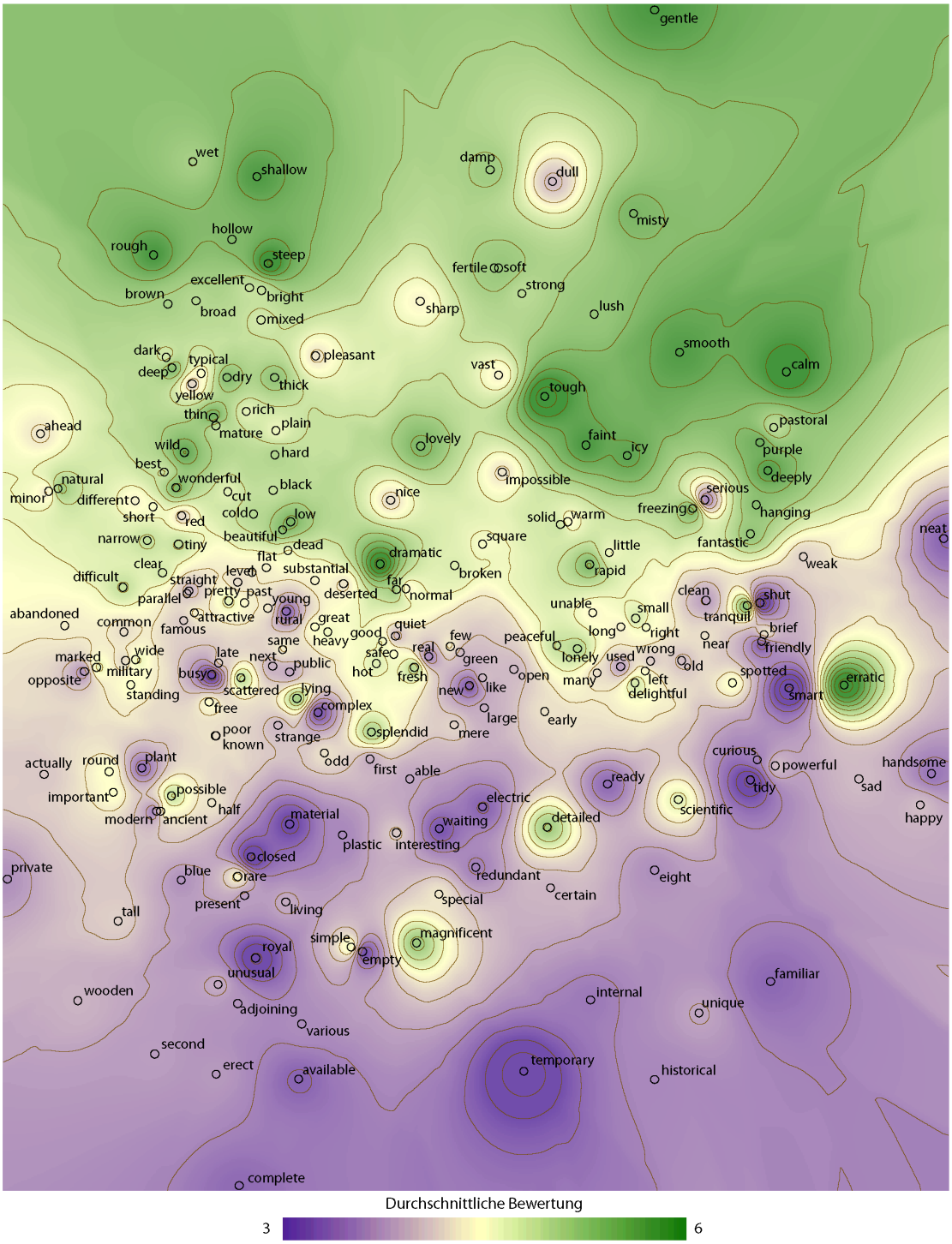


Abb. 6.10: Ergebnis der multidimensionalen Skalierung mit Adjektiven.

In der Abb. 6.10 ist das Ergebnis der multidimensionalen Skalierung mit 200 häufigen Adjektiven zu sehen. Es gibt wiederum ein Bewertungsgefälle zwischen naturräumlich

konnotierten Begriffen wie *shallow* oder *misty* und urbanen Begriffen wie *electric* oder *busy*. Im oberen Bereich sind Adjektive angesiedelt, die man tendenziell in Beschreibungen des natürlichen Raumes erwartet. Unten findet man mehrheitlich Adjektive des urbanen Raumes. Es bildet sich bei den Adjektiven aber kein derart deutlicher Kontinentalrand wie in Abb. 6.9 aus. Vielmehr gleicht die Bewertungstopografie einer seichten Küste mit vorgelagerten Inseln. Dies ist ein Hinweis auf die geringere Bewertungsspannweite der Adjektive. Trotzdem lässt sich in der Darstellung eine Systematik erkennen.

6.1.3 Ortsabhängigkeit der Bewertung einzelner Begriffe

Die sehr aktiven Fotografen auf Geograph beschränken sich meist auf räumlich begrenzte Gebiete, wie in Abschnitt 3.1.1 auf Seite 23 festgestellt wurde. In einem weiteren Schritt wurde die durchschnittliche Bewertung aller Begriffe, die diese User in die Geograph-Kommentare geschrieben haben, für jeden einzelnen der 35 Nutzer berechnet. Zusätzlich wurde für jeden der 35 aktivsten Fotografen der geografische Schwerpunkt der zugehörigen Fotografien berechnet. Die Verknüpfung der Schwerpunkte und der Bewertung einzelner Begriffe enthüllt, dass die durchschnittlichen Bewertungen einem bestimmten Muster folgen. Je nördlicher der Schwerpunkt des Users liegt, desto besser ist in der Tendenz die durchschnittliche Bewertung desselben Begriffs. Der Begriff *road* hat in südlichen Regionen teilweise einen Wert von weniger als 4 – im Norden hingegen über 5. Allerdings wurde gerade beim Begriff *road* wahrscheinlich typischerweise nicht die in der Legende beschriebene Strasse, sondern die umgebende Landschaft bewertet. Das beschriebene Muster ist aber nicht nur bei diesem Begriff zu finden, sondern auch bei zahlreichen anderen, meist naturräumlichen Begriffen. Besonders unterschiedlich wird der Begriff *ridge* (Grat) bewertet, dessen durchschnittliche Bewertungen von 2,8 bis 7,3 reichen, wobei die positiveren Bewertungen vorwiegend im Norden zu finden sind. Der Begriff wird, wie in der Darstellung 6.11 zu sehen ist, im ganzen Untersuchungsgebiet verwendet. Weitere Begriffe mit diesem Verhalten sind zum Beispiel *valley*, *water*, *river* und *steep*.

Umgekehrt verhält es sich hingegen beim Wort *industrial*, das, wenn es im Norden Grossbritanniens verwendet wird, besonders negativ bewertet wird, in den südlicheren Gebieten hingegen weniger negativ. Daraus kann man schliessen, dass auf das Vorkommen industrieller Einrichtungen in den generell positiv bewerteten Landschaften des Nordens speziell sensibel reagiert wird. Dasselbe Phänomen kann auch für die Begriffe *motorway* und *shop* beobachtet werden, die auf Bildern von Fotografen des Nordens negativer bewertet werden als bei Fotografen mit einem südlicher liegenden Schwerpunkt.

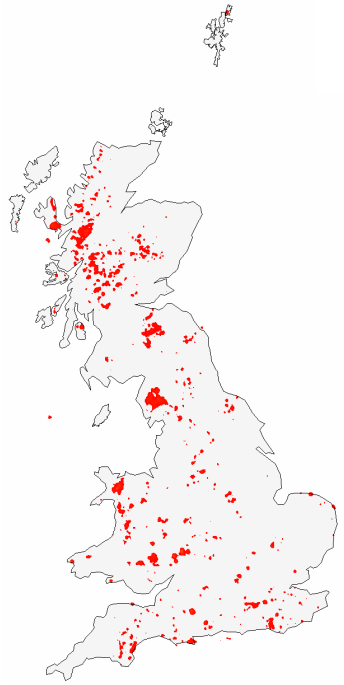


Abb. 6.11: Verteilung von Fotos mit dem Begriff «ridge» (Grat) in der Legende.

Die Beantwortung der Forschungsfrage 1 kann folgendermassen erweitert werden: Tendenziell werden also Begriffe des urbanen Raums (generell negativ bewertet) in nördlichen Gebieten besonders negativ bewertet und im Süden etwas weniger negativ, währenddem das Verhältnis bei naturräumlichen Begriffen (generell positiv bewertet) genau umgekehrt ist.

6.2 Schätzung der Landschaftsbewertung

6.2.1 Vorbemerkungen

Nachfolgend werden die Resultate der diversen Schätzmethoden wiedergegeben. Es wurde jeweils eine Auswahl des Validierungssets getroffen und für welche die Schätzwerte berechnet wurden. Gleichzeitig sind auch die wirklichen Landschaftsbewertungen bekannt, welche dann in Bezug zur Schätzung gesetzt werden. Die Güte der einzelnen Schätzungen wird durch das Bestimmtheitsmass r^2 angegeben, das zeigt, welcher Anteil der Varianz durch das Sprachmodell erklärt werden kann.

6.2.2 Unigramm-Schätzung

In diesem Abschnitt werden die Resultate der insgesamt 500 Kombinationen der Unigrammschätzung dargelegt. Die grosse Zahl der Kombinationen erlaubt es, die Effekte der einzelnen Parameter zu beschreiben.

Gewichtungsparameter

Im Abschnitt 5.5.3 wurde die Gewichtungsfunktion der Begriffswerte beschrieben. Um zu evaluieren, welcher Exponent m die besten Resultate erzielt, wurden fünf verschiedene Werte getestet. Die Ergebnisse, die dem Durchschnitt über die anderen Parameter entsprechen, sind in der Tabelle 6.2 dargestellt.

Exponent m	0.1	0.2	0.3	0.4	0.5
r^2	44.07 %	44.16 %	44.24 %	44.31 %	44.39 %

Tabelle 6.2: Einfluss des Exponenten m auf r^2

Insgesamt ist der Einfluss der Wahl des Exponenten klein. Wie sich zeigt, steigt das Bestimmtheitsmass mit grösser werdendem Exponenten der Gewichtungsfunktion sehr leicht. Für die folgenden Auswertungen werden deshalb immer die Berechnungen mit einem Exponenten von 0.5 verwendet.

Einfluss der maximalen Varianz

In der Abb. 6.12 sind die Bestimmtheitsmasse in Abhängigkeit von der maximalen Varianz der Stimmen dargestellt. Der Einfluss aller anderen Parameter wurde gemittelt. Die maximale Varianz von 2.4 bedeutet, dass die Hälfte der Bilder einbezogen wird

(siehe Abschnitt 3.2), bei einer maximalen Stimmvarianz von 50 werden alle Bilder berücksichtigt. In den grünlichen Tönen sind die Resultate für die transformierten Werte (gemäss Berechnung in Abschnitt 5.5.2) dargestellt. Bei diesem Vorgehen wurden die Begriffswerte vor der Schätzung transformiert. Die erdigen Farben zeigen die Ergebnisse der ursprünglichen Durchschnittswerte des Unigrammlexikons.

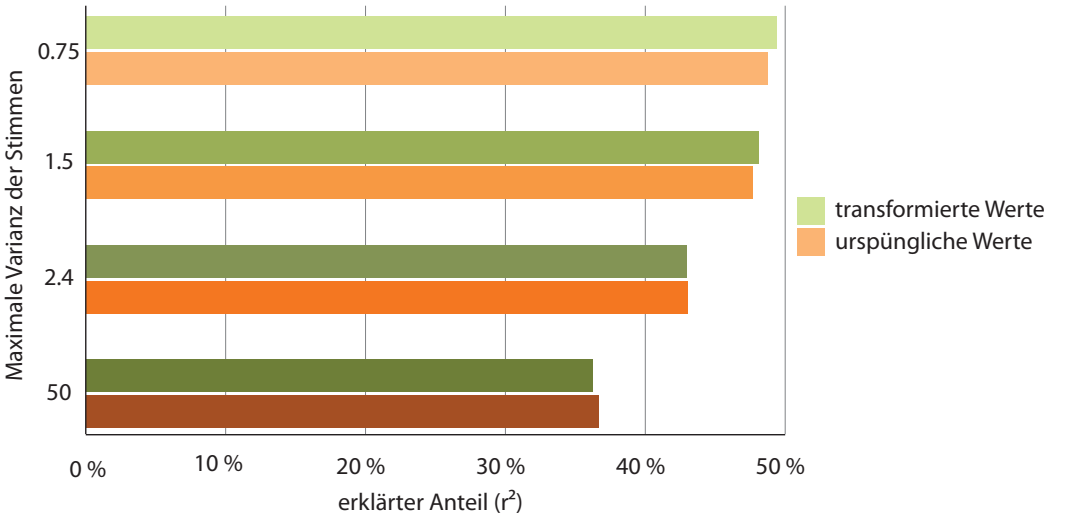


Abb. 6.12: Einfluss der maximalen Varianz auf das Bestimmtheitsmass.

Es zeigt sich, dass die Güte des Modells abnimmt, je grösser die zugelassene Varianz der Stimmen im Validierungsset ist. Bei einer maximalen Varianz von 0.75 kann fast 50 % der Stimmenvergabe erklärt werden. Wenn alle Bilder miteinbezogen werden, sinkt dieser Wert auf 36 %. Ferner ist auch ersichtlich, dass der erklärte Anteil bei kleiner Varianz bei der Schätzung mit den transformierten Werten leicht besser ist als mit den ursprünglichen Werten – wenn alle Bilder einbezogen werden, erzielen dagegen die ursprünglichen Werte ein leicht besseres Resultat. Da die transformierten Werte insgesamt ein etwas besseres Ergebnis ermöglichen, werden in den folgenden Betrachtungen diese Werte benutzt.

Insgesamt zeigt die Abhängigkeit von der Stimmvarianz, dass sich eine Landschaftsbewertung besser schätzen lässt, wenn sich die ScenicOrNot-Bewerter relativ einig waren, also bei kleiner Varianz.

Einfluss der Anzahl Worte im Kommentar

Ein weiterer Parameter, dessen Einfluss auf die Ergebnisse getestet wurde, ist die minimale Anzahl Worte, die ein Kommentar aufweisen muss, um in die Berechnung einzufliessen. Das Resultat ist in der Abb. 6.13 dargestellt, zusätzlich werden die Bestimmtheitsmasse

nach den Varianzen der Stimmen aufgeschlüsselt. Der Effekt des Mindestvorkommens der Begriffe im Trainingsset wird hier noch nicht berücksichtigt.

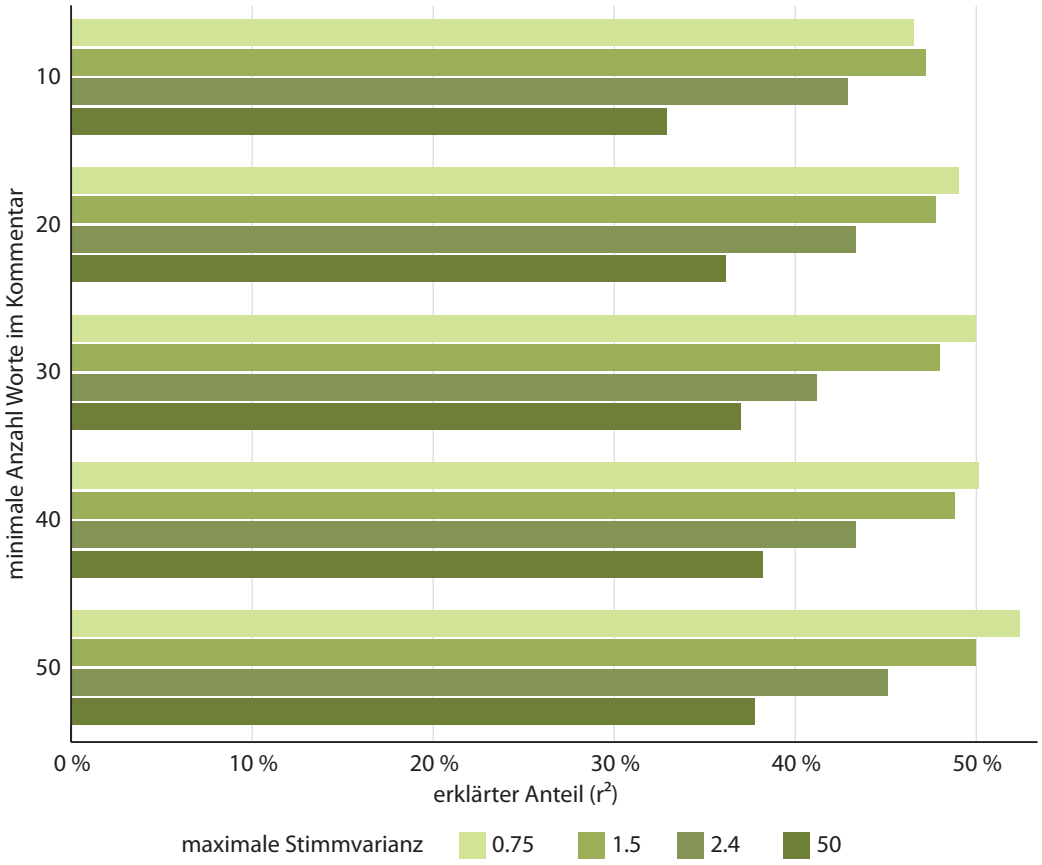


Abb. 6.13: Einfluss der Anzahl Worte im Kommentar.

Es sind zwei deutliche Trends erkennbar: Einerseits nimmt, wie bereits oben beschrieben, die Güte der Schätzung mit zunehmender Varianz ab; andererseits verbessert sich die Schätzung mit zunehmender Anzahl Worte, die im Kommentar vorhanden sein müssen. Das beste Bestimmtheitsmass, das so erreicht wird, beträgt 0.524. Mit diesem Modell ist demnach mehr als die Hälfte der Varianz der Schönheitsbewertungen erklärbar.

Je mehr Worte ein Kommentar von Geograph enthält, desto präziser lässt sich die Ästhetik der beschriebenen Landschaft schätzen. Dies ist nicht selbstverständlich, denn gerade bei längeren Kommentaren besteht auch die Möglichkeit, dass mehr als nur die dargestellte Landschaft beschrieben wird, etwa die Geschichte des fotografierten Ortes. Längere Kommentare, und das ist der Wermutstropfen dieser Erkenntnis, sind in den Geograph-Daten deutlich seltener. Wenn die Bilderauswahl mit einer maximalen Stimmvarianz von 0.75 und mindestens 50 Worten eingeschränkt wird, gibt es nur gut 500 Bilder des Validierungssets, die diesen Kriterien entsprechen. Rigide Anforderungen verbessern die Schätzung, schliessen aber viele Bilder im Vornherein aus.

Einfluss des Mindestvorkommens im Trainingsset

Bei der Generierung des Begriffslexikons für das Trainingsset wurde gezählt, wie oft jeder erfasste Begriff in den Kommentaren vorkam. Wie die Ausführungen auf Seite 43 dokumentiert haben, nimmt die Stabilität der Wortbewertungen mit zunehmendem Mindestvorkommen zu. Auf dieser Beobachtung fusst die Vermutung, eine Schätzung lasse sich mit dem Ausschluss selten vorkommender Begriffe verbessern, weil deren Wert unsicher ist. Um dies zu überprüfen, wurde die Schätzung mit fünf verschiedenen Werten für das Mindestvorkommen durchgeführt. In der Abb. 6.14 sind die Befunde dieser Berechnungen dargestellt, auch hier aufgeschlüsselt nach maximaler Varianz.

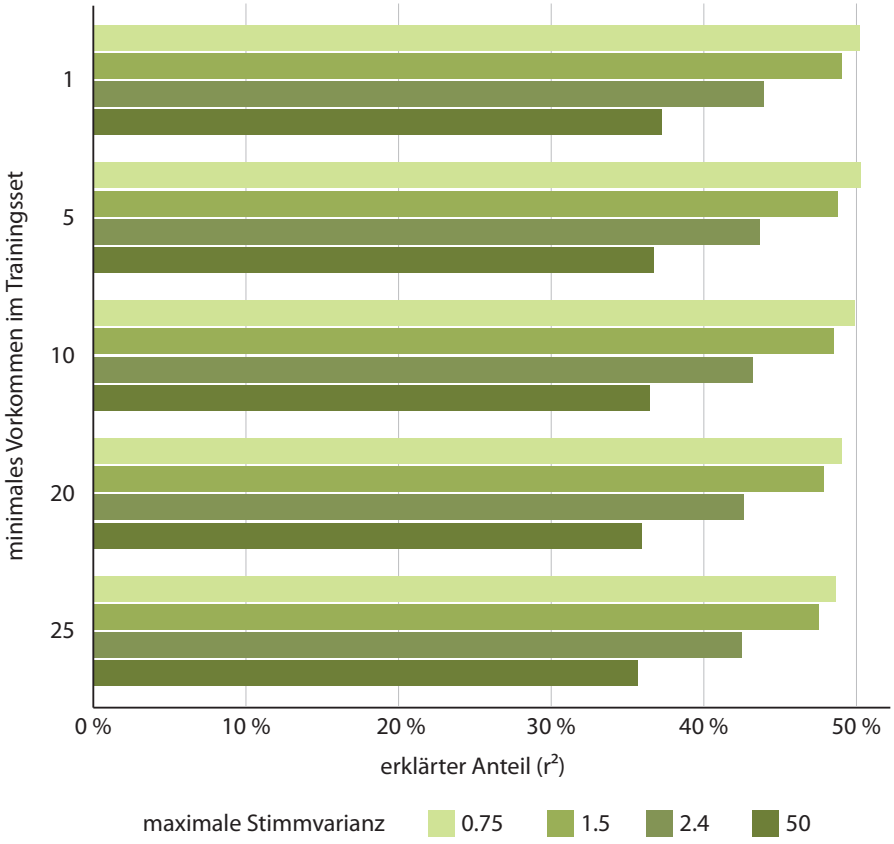


Abb. 6.14: Einfluss des Mindestvorkommens eines Begriffs im Trainingsset bei der Unigramm-Schätzung.

Entgegen der Vermutung sinkt die Qualität der Schätzung mit zunehmendem Mindestvorkommen. Die tiefen Mindestvorkommen erreichen ein r^2 von 0.503 (mindestens 5 Mal) bzw. 0.502 (> 0). Wenn die Anforderungen an das Trainingsset steigen, nimmt das Bestimmtheitsmass leicht ab und fällt unter 0.5. Offensichtlich dominiert der Vorteil der grösseren Trefferquote beim Nachschlagen der Begriffswerte über die grössere Sicherheit der Bewertung der oft vorkommenden Begriffe. Insgesamt ist die Auswirkung des mini-

malen Vorkommens aber eher bescheiden. Wesentlich grösser und erneut sehr ins Auge stechend ist der Effekt der maximalen Varianz.

Kombination von Anzahl Wörtern und Mindestvorkommen im Trainingsset

Um das bestmögliche Resultat zu evaluieren, müssen die Parameter Mindestvorkommen und Mindestanzahl Worte kombiniert werden. Dies lässt sich aus den oben vorgestellten Resultaten schliessen. In der Abb. 6.15 ist das Ergebnis dieser Kombination dargestellt. Hier wurden nur die kleinsten Varianzen (<0.75, dunkle Töne) und alle Varianzen (helle Töne) dargestellt. Es wird der erklärte Anteil in Abhängigkeit der Mindestanzahl Wörter im Kommentar abgebildet, zusätzlich aufgeschlüsselt nach dem Mindestvorkommen der Begriffe im Trainingsset.

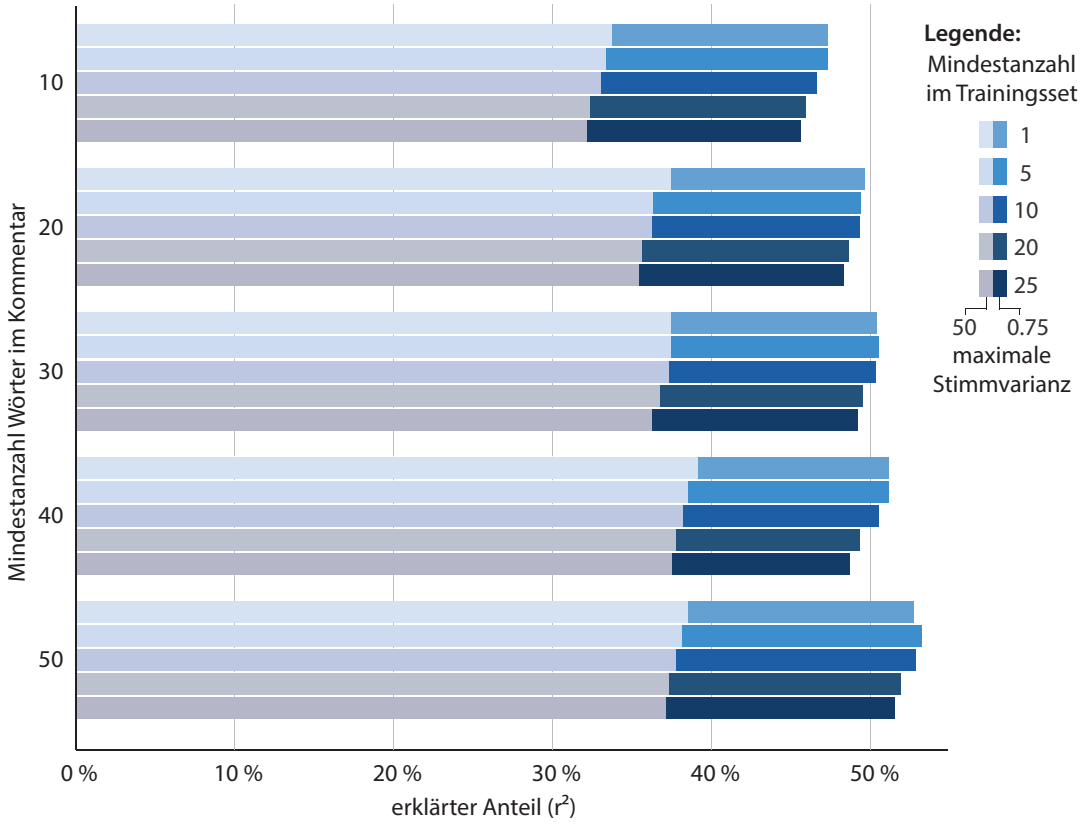


Abb. 6.15: Einfluss des Mindestvorkommens und der Kommentarlänge auf die Güte der Unigramm-Schätzung. Der helle linke Teil der Balken zeigt die Resultate für eine maximale Stimmvarianz von 50; der dunklere die Resultate für eine eingeschränkte Stimmvarianz von 0.75.

Die beste Schätzung gelingt mit langen Kommentaren und geringen Anforderungen an das Mindestvorkommen im Begriffslexikon. Bei einem Mindestvorkommen von 5 und

mindestens 50 Wörtern im Kommentar erreicht r^2 den Höchstwert von 0.532, wenn die maximale Varianz der ScenicOrNot-Bewertungen auf 0.75 limitiert wird. Bei den übrigen Mindestlängen der Kommentare wird jeweils bei einem Mindestvorkommen von 1 im Trainingsset das beste Resultat erzielt. Bei einer Mindestlänge von 10 Worten erreicht r^2 immer noch einen Wert von 0.473 wenn alle Worte des Trainingssets miteinbezogen werden. Neben den bisher erwähnten Werten für eine maximale Varianz von 0.75 sind in den helleren Tönen auch die Bestimmtheitsmasse für unbeschränkte Varianzen aufgeführt. Das Grundmuster der oben beschriebenen Abhängigkeiten bleibt auch hier erhalten, aber die Bestimmtheitsmasse sinken um 12 bis 15 Prozentpunkte. r^2 beträgt bei einem jeweiligen Mindestvorkommen von 1 0.337 (min. 10 Worte im Kommentar) bzw. 0.385 (min. 50 Worte im Kommentar). Bei einer maximalen Stimmvarianz von 2.41, was die Hälfte der ScenicOrNot-Bilder umfasst, wird immerhin ein r^2 von 0.45 erzielt.

Neben der Übereinstimmung der Noten kann zur Evaluation der Schätzung auch analysiert werden, wie präzis die Rangreihenfolge einer Anzahl Bilder reproduziert werden kann. *Shafer und Mietz* (1970) haben ihr Modell so überprüft. Mit denselben Vorgaben, die den Höchstwert von 0.532 erzielt haben, kann ein Rangkorrelationskoeffizient (Spearmans Rho) von 0.648 erreicht werden.

Zusammenfassend lässt sich folglich festhalten, dass die Schätzung besser wird je mehr Worte in diese einfliessen – also bei langen Kommentaren mit geringen Anforderungen an die Häufigkeit im Trainingsset. Die Annahme, die Schätzung verbessere sich mit immer längeren Kommentaren stetig weiter, ist aber falsch. Bei mindestens 60 Worten verbessert sich die Schätzung noch minimal auf 0.538, aber schon bei mindestens 70 Worten verschlechtert sie sich auf 0.527 und bei mindestens 80 Worten fällt sie auf 0.498. Bei den längeren Texten steigt die Tendenz, dass viele Hintergrundinformationen über die Landschaft wiedergegeben werden, die nicht das direkt sichtbare beschreiben.

Trainingsset mit geringer Varianz

Bei der Bildung des Trainingssets kann die Auswahl der Bilder dahingehend eingeschränkt werden, dass nur solche mit relativ geringer Varianz berücksichtigt werden. Die Idee dahinter besteht darin, nur Kommentare von Bildern zu analysieren, bei deren Beurteilung sich die ScenicOrNot-Bewerter relativ einig waren, um damit eine sichere Benotung der einzelnen Begriffe zu erreichen. Als Grenzwert für die maximale Varianz im Rating wurde hier 3 verwendet. Wenn dieselben Vorgaben, die oben zum besten Resultat von 0.532 geführt haben (Mindestvorkommen von 5 und mindestens 50 Wörter im Kommentar), verwendet werden, wird ein etwas schlechteres Ergebnis von 0.522 erreicht, also ein Prozentpunkt weniger. Der Grund dafür dürfte abermals in der kleineren Anzahl Begriffe

des Trainingssets, die zum Nachschlagen zur Verfügung stehen, liegen. Nur Bilder mit geringer Varianz für die Erstellung des Trainingssets zu nutzen führt demzufolge nicht zu besseren Resultaten.

Verwendung des nutzergetrennten Trainingssets

Neben dem bisher verwendeten Unigrammlexikon, das auf einer zufällig ausgewählten Hälfte der ScenicOrNot-Daten basiert, wurde auch ein nutzergetrenntes Trainingsset erstellt. Das Validierungsset enthält in diesem Fall keine Bilder von Nutzern, deren Bilder und Kommentare in die Bildung des Trainingssets eingeflossen sind. Dadurch haben mögliche Eigenheiten eines Nutzers in der Wortwahl keine Auswirkung, weil das Trainings- und das Validierungsset vollständig getrennt sind. Die Ausführungen auf der Seite 46 haben belegt, dass die Stabilität zwischen den beiden nutzergetrennten Lexika etwas geringer ist. Dies zeigt sich auch bei der Schätzung der Landschaftsbewertungen mit einem nutzergetrennten Lexikon: Es wird bei Anwendung der oben gefundenen optimalen Einstellungen (Mindestvorkommen von 5 und mindestens 50 Wörter im Kommentar) ein r^2 von 0.495 erreicht. Dies weist auf eine gewisse Verbindung zwischen Trainings- und Validierungsset hin, wenn derselbe Nutzer in beiden vorkommen kann.

Toponymerkennung

Bei den bisherigen Ausführungen wurden die Ortsnamen stets aus den Kommentaren herausgefiltert und beim Nachschlagen nicht berücksichtigt. Wenn diese Vorgaben nicht beachtet werden, kann das Bestimmtheitsmass weiter gesteigert werden: Bei einem Mindestvorkommen von 5 mal im Trainingsset und 50 Wörtern im Kommentar wird ein r^2 von 0.549 erzielt. Wenn die Toponyme in die Schätzung einbezogen werden, kann die Schätzung verbessert werden. Erstaunlich ist dies nicht, denn abgesehen von Ambiguitäten bieten die Ortsbezeichnungen eine sichere Bewertung über die Ästhetik eines Ortes und können darum zu einer Verfeinerung der Schätzung beitragen.

Synonymsuche

Es ist durchaus möglich, für ein Wort keine Entsprechung im Lexikon zu finden, insbesondere wenn die Anforderung an das Mindestvorkommen bei der Schätzung hoch ist. Mit einem Synonymwörterbuch besteht die Möglichkeit, ein sinngemässes Wort zu suchen, das im Lexikon existiert und dann dessen Wert zu verwenden.

Bei einer maximalen Stimmvarianz von 0.75, mindestens 50 Worten im Kommentar und mindestens 5 Treffern im Trainingsset verschlechtert sich r^2 von 0.532 mit der

Synonymsuche auf 0.525. Auch wenn die Anforderungen des Mindestvorkommens im Trainingsset auf 25 erhöht werden, verschlechtert sich r^2 bei sonst gleichen Bedingungen von 0.515 auf 0.499. Mit dem vorliegenden Verfahren ergeben sich durch die Synonymsuche also keine Vorteile für die Landschaftsschätzung. Es bleibt aber anzumerken, dass die automatische, unüberwachte Auswahl eines passenden Synonyms eine schwierige Aufgabe ist.

Einbezug des gesamten Validierungssets

In den bisherigen Analysen wurde jeweils ein zufällig ausgewähltes Subset des Validierungssets ausgewählt, damit die Rechenzeit in akzeptablem Rahmen bleibt. Um zu untersuchen, wie sich diese Verkleinerung des Validierungssets auf das Resultat auswirkt, wurde auch eine Schätzung mit allen Bildern, die folgende Bedingungen erfüllen, durchgeführt: Unbeschränkte Stimmvarianz und mindestens 10 Wörter im Kommentar. Dies trifft auf 55'500 Bilder zu. Bei einer Unigramm-Schätzung mit einem Mindestvorkommen von 5 im Lexikon wird damit ein r^2 von 0.344 erreicht. Dies ist ein ähnliches Resultat wie es mit dem Subset (0.341) erreicht wird. Die Einschränkung der Grösse des Validierungssets ist folglich zulässig.

6.2.3 Lokale Unigramm-Schätzung

Die lokale Unigramm-Schätzung verwendet regional variierende Werte für dieselben Begriffe und stellt damit eine Verfeinerung der Unigramm-Variante dar. Zur Schätzung wird jeweils der nächste Schwerpunkt eines sehr aktiven Geograph-Fotografen gesucht und dann die Schönheitswerte, die auf dessen Fotografien basieren, verwendet.

Parameter	**Werte**				
Minimale Anzahl Worte im Kommentar	10	10	10	25	25
Minimale Anzahl im Trainingsset	1	2	2	2	10
maximale Varianz	1.5	1	1.5	1.5	1.5
r^2	0.248	0.245	0.248	0.225	0.172

Tabelle 6.3: Resultate des Bestimmtheitsmasses mit 5 Parameterkombinationen für die Schätzung mit lokalen Unigrammen

In der Tab. 6.3 sind die Resultate der Schätzung der Bildbewertung mit lokalen Unigrammen wiedergegeben. Entgegen der Vermutung sind die Korrelationswerte sehr tief. r^2 nimmt leicht ab, wenn die Mindestanzahl der Anzahl Wörter bei der Bildauswahl erhöht wird. Den grösseren Einfluss auf das Bestimmtheitsmass hat aber die Anzahl, wie oft ein

Begriff im Trainingsset gefunden werden musste, wie das Resultat für die Mindestzahl 10 zeigt. Dies dürfte darauf zurückzuführen sein, dass die zur Schätzung zur Verfügung stehenden Begriffe durch diese Einschränkung rapide abnehmen. Bei weniger rigiden Einschränkungen stehen mehr Begriffe aus dem Trainingslexikon zur Verfügung, was die Schätzung leicht verbessert. Der Grund für die tiefen Korrelationswerte liegt aber letztlich wohl in der Unsicherheit der Durchschnittswerte der Begriffe. Viele Begriffe wurden pro User nur ein- bis zweimal gefunden, was bedeutet, dass der Schönheitswert von eben so wenigen Bildern abhängt und dadurch sehr anfällig auf Ausreisser ist.

Möglicherweise wären die Resultate besser, wenn wesentlich mehr Kommentare für die Bildung des Trainingssets zur Verfügung stehen würden. Dann hätten die einzelnen Begriffe eine breiter abgestützte Bewertungsbasis, was sich positiv auf ihre Stabilität auswirken würde. Mit der verwendeten Datenbasis und dem daraus resultierenden Trainingsset scheint es aber nicht lohnend, Landschaftsbewertungen auf Basis der lokalen Unigramme weiter zu verfolgen.

6.2.4 Kookkurrenz-Schätzung

Auch für die Schätzung mit dem Kookkurrenzlexikon wurde eine Reihe von Parameterkombinationen getestet. Weil sich bei den Analysen mit den Unigrammen früh gezeigt hat, dass der Gewichtungsexponent $m = 0.5$ am sinnvollsten ist, wurde nur dieser für die Berechnungen der Kookkurrenz-Schätzungen verwendet. Nachfolgend werden die damit erzielten Ergebnisse präsentiert – der Aufbau gestaltet sich ähnlich wie bei den Unigrammen.

Einfluss der maximalen Stimmvarianz

maximale Varianz	0.75	1.5	2.4	50
r^2	40.14 %	37.97 %	33.37 %	29.5 %

Tabelle 6.4: Einfluss der maximalen Varianz auf r^2

Die Bestimmtheitsmasse der Kookkurrenzanalyse sind ebenfalls in grossem Mass von der maximal zugelassenen Varianz im Validierungsset abhängig. Die durchschnittlichen Werte sind in der Tabelle 6.4 wiedergegeben. Wiederum nimmt die Güte der Schätzung mit grösser werdender Stimmvarianz ab, die Werte sind aber durchgängig geringer als bei der Unigrammschätzung, denn sie liegen durchschnittlich gut 9 Prozentpunkte tiefer.

Einfluss der Anzahl Worte im Kommentar

Auch bei der Bestimmung der Landschaftsästhetik mit dem Kookkurrenzlexikon hat die Länge der Kommentare eine Wirkung. Die Unterschiede sind in der Abb. 6.16 dargestellt. Bei Betrachtung der kleinsten Varianz, 0.75, zeigt sich grundsätzlich dasselbe Bild wie bei den Unigrammen: Je länger die Kommentare sind, desto besser die Schätzung, wobei die Unterschiede geringer als bei den Unigrammen ausfallen.

Bei höheren zugelassenen Varianzen sinkt die Güte der Schätzung – auch das deckt sich mit den Beobachtungen bei den Unigrammen. Allerdings nimmt r^2 so stark ab, dass die Güte der Schätzung mit zunehmender Anzahl Worte im Kommentar abnimmt. So wird bei mindestens 10 Wörtern und einer beliebigen Stimmvarianz ein r^2 von 0.306 erreicht, bei mindestens 50 Wörtern aber nur noch 0.274. Es sind hier also zwei gegenläufige Trends zu beobachten: Mit zunehmender Anzahl Wörter im Kommentar steigt r^2 bei kleinen Varianzen, bei grossen hingegen sinkt r^2.

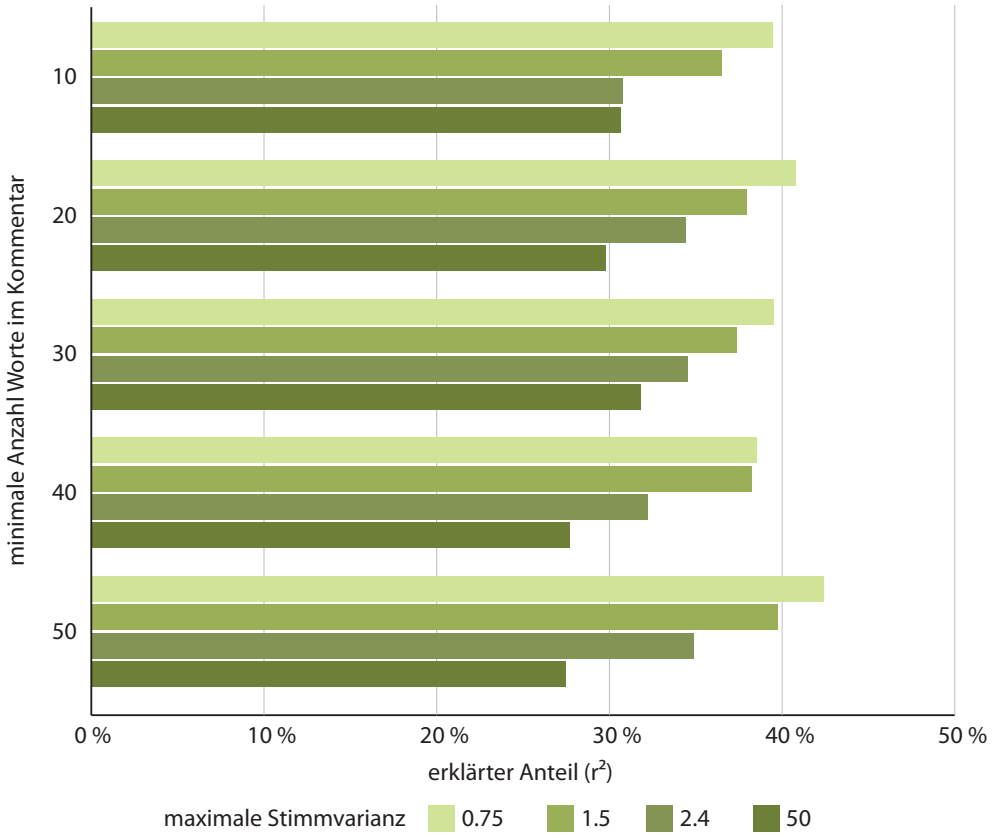


Abb. 6.16: Einfluss der Anzahl Worte im Kommentar bei der Kookkurrenzschätzung.

Einfluss des Mindestvorkommens im Trainingsset

Bei der Erstellung des Kookkurrenz-Lexikons wurde das Vorkommen der Begriffspaare im Trainingsset gezählt. Die Abb. 6.17 zeigt das Bestimmtheitsmass in Abhängigkeit des minimalen Vorkommens im Trainingsset. Nicht überraschend ist das beste Abschneiden der kleinsten maximalen Stimmvarianz. Im Gegensatz zur Unigramm-Methode wird hier nicht mit dem geringsten minimalen Vorkommen der beste Wert erreicht, sondern bei einer Mindestzahl der Begriffspaare im Trainingsset von 5. Bei einer maximalen Varianz von 0.75 beträgt r^2 in diesem Fall 0.42. Bei der Steigerung des Mindestvorkommens von 1 auf 5 halbiert sich die zur Verfügung stehende Zahl der Begriffspaare, trotzdem verbessert sich die Schätzung. Dies kann als Indiz dafür gesehen werden, dass die selten vorkommenden Begriffspaare in ihrer Bewertung recht unsicher sind.

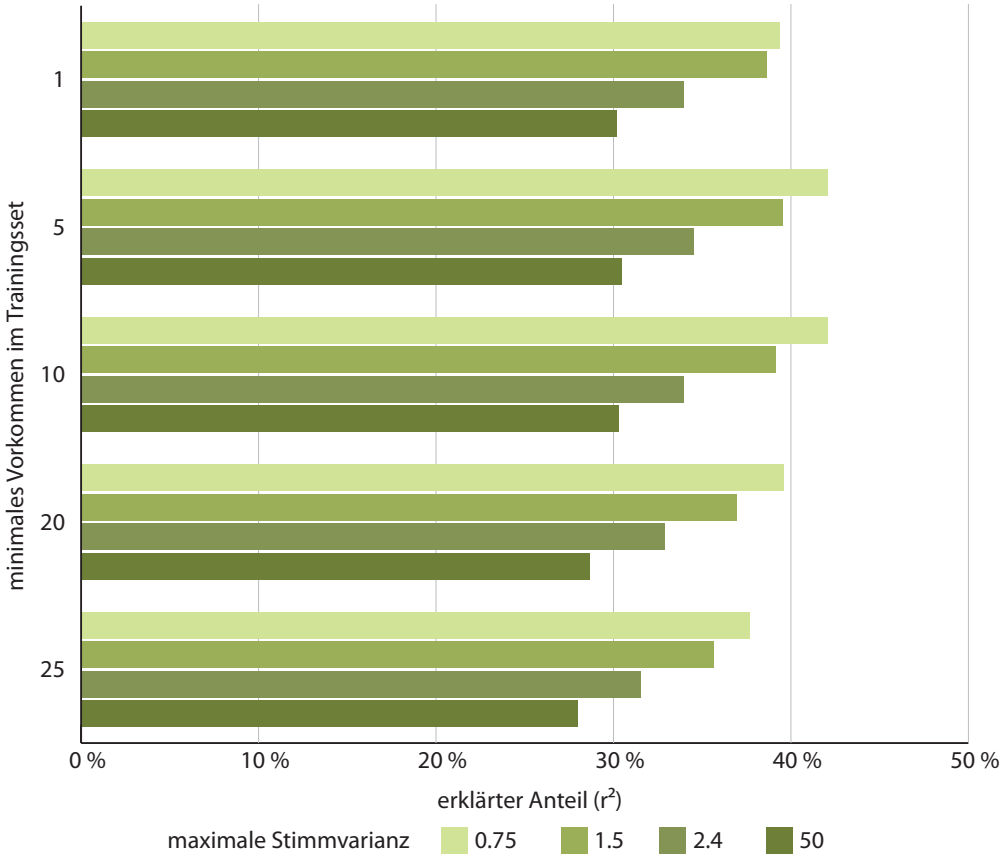


Abb. 6.17: Einfluss des Mindestvorkommens im Trainingsset bei der Kookkurrenz-Schätzung.

Mit weiter steigendem Mindestvorkommen sinkt die Qualität der Schätzung wieder leicht – bei grösserer Einschränkung der maximalen Varianz stärker als bei diesbezüglich lascheren Vorgaben. Wenn die Begriffspaare häufig vorkommen, ist zwar ihre Bewertung stabil, weil

sie auf zahlreichen Stimmen basiert, aber gleichzeitig stehen weniger Begriffspaare zur Schätzung zur Verfügung. Dies ist der Grund für die schlechter werdende Schätzung.

Es scheint also ein Optimum zwischen der Sicherheit der Begriffsbewertung und der Zahl der zur Verfügung stehenden Begriffspaare zu geben.

Kombination von Kommentarlänge und Mindestvorkommen

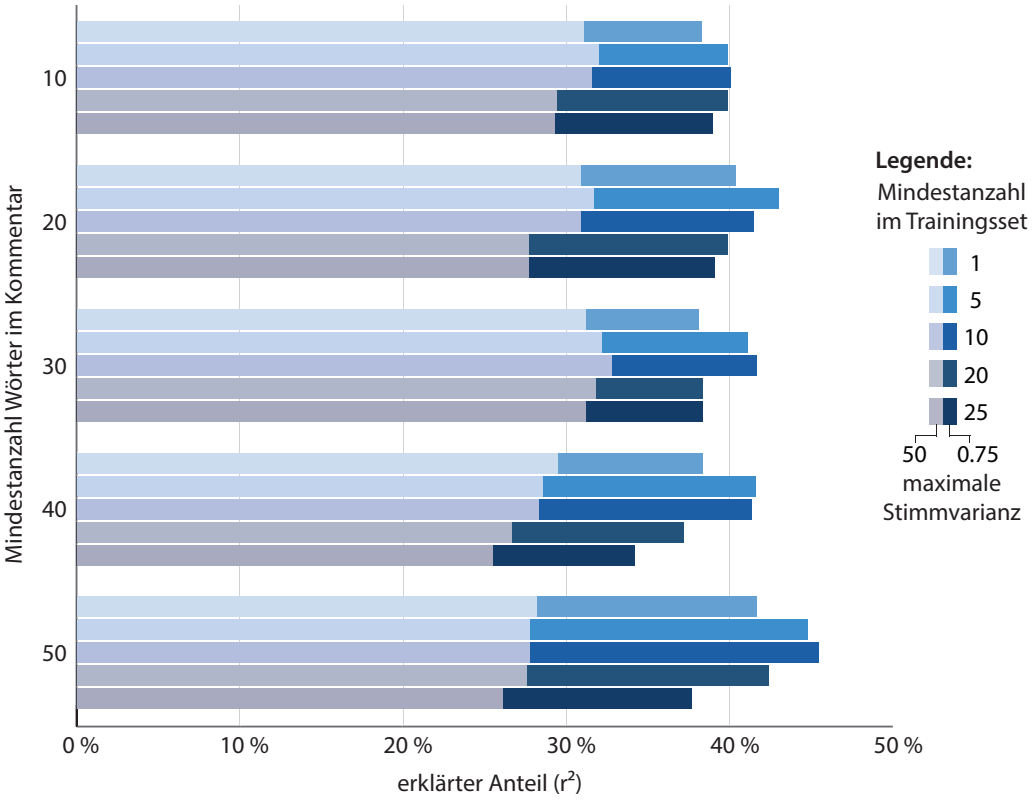


Abb. 6.18: Einfluss des Mindestvorkommens und der Kommentarlänge auf die Güte der Kookkurrenz-Schätzung. Der helle linke Teil der Balken zeigt die Resultate für eine maximale Stimmvarianz von 50; der dunklere die Resultate für eine eingeschränkte Stimmvarianz von 0.75.

In der Abb. 6.18 sind die detaillierten Ergebnisse des Zusammenhangs zwischen Mindestlänge des Kommentars und der Mindestanzahl Treffer im Kookkurrenzlexikon dargestellt. Die dunklen Blautöne stellen wiederum eine maximale Stimmvarianz von 0.75 dar und in den helleren Blautönen sind die Resultate für unbeschränkte Stimmvarianzen veranschaulicht. Wie bereits aufgrund der oben präsentierten Resultate vermutet werden kann, sind die besten Resultate bei langen Kommentaren und einem mittleren Mindestvorkommen im Trainingsset zu finden. Bei mindestens 10 Treffern und mindestens 50 Worten im

Kommentar wird ein Bestimmtheitsmass von 0.455 erreicht, was den Bestwert darstellt. Dieser Wert wird auch hier bei einer maximalen Stimmvarianz von 0.75 erreicht.

Wenn die Varianzen bei der Bilderauswahl nicht eingeschränkt werden, wird das beste Resultat nicht mehr bei besonders langen Kommentaren erreicht, sondern bei mindestens 30 Worten im Kommentar. Mit dem Kookkurrenzverfahren wird so ein r^2 von 0.328 erreicht.

Insgesamt bleiben die Bestimmtheitsmasse deutlich hinter dem Unigrammverfahren zurück.

6.2.5 Schätzung mit häufigen Begriffen

Die 560 Begriffe, deren gemeinsames Vorkommen in der Kookkurrenz-Schätzung geprüft wurde, können auch im Sinne der Unigramme für eine Schätzung verwendet werden. Im Kern bedeutet dies, dass ein stark verkleinertes Unigrammlexikon zum Nachschlagen verwendet wird und die Schätzung nur auf diesen Begriffen basiert. Da alle diese Begriffe in den Geograph-Begleittexten häufig vorkommen, macht es keinen Sinn, das Mindestvorkommen im Trainingsset zu untersuchen. In der Abb. 6.19 sind daher nur 20 Kombinationen von vier verschiedenen maximalen Stimmvarianzen und fünf minimalen Kommentarlängen wiedergegeben.

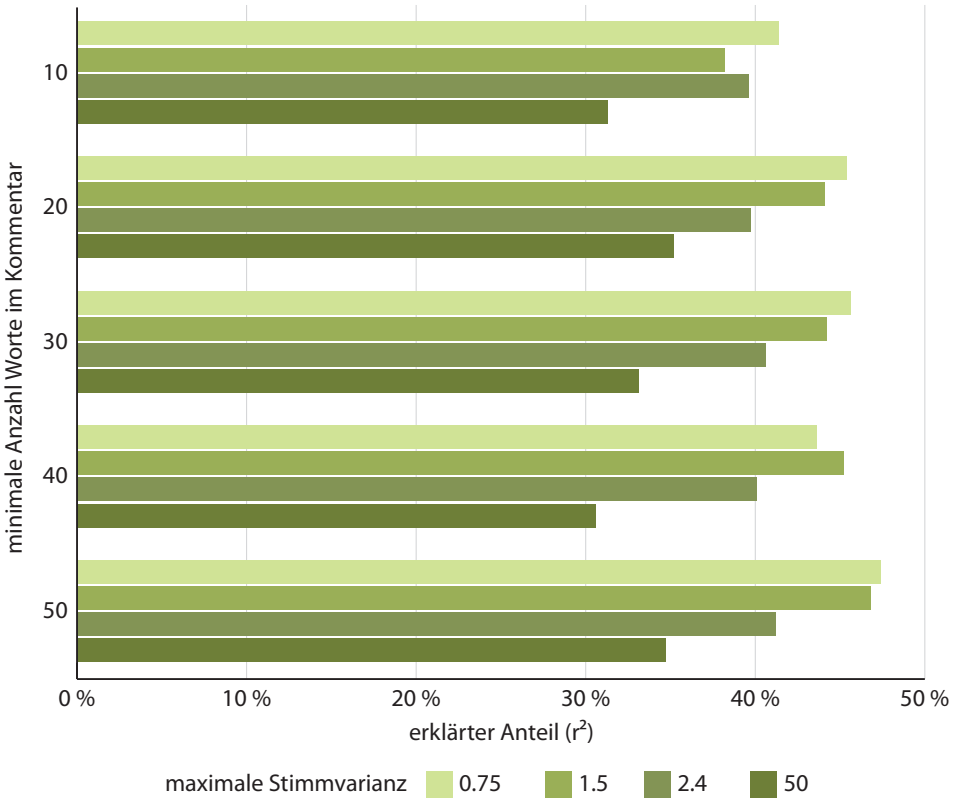


Abb. 6.19: Resultate der Schätzung nur mit sehr häufigen Begriffen.

Generell wird die Güte der Schätzung kleiner, wenn die maximal zugelassene Stimmvarianz grösser wird. Es gibt aber Ausnahmen von diesem Trend, so etwa bei den kurzen Kommentaren, wo eine maximale Stimmvarianz von 2.41 schlechter abschneidet als 1.5. Ausserdem verbessert sich auch hier die Schätzung bei länger werdenden Kommentaren. Bei mindestens 50 Worten im Kommentar und einer maximalen Varianz von 0.75 wird in diesem Verfahren ein maximales r^2 von 0.474 erreicht, also nur knapp 6 Prozentpunkte weniger als bei der feineren Unigrammschätzung und knapp zwei Prozentpunkte mehr als der beste Wert der Kookkurrenz-Schätzung.

Die Reduktion des Bewertungslexikons auf eine überschaubare Anzahl sehr häufiger Begriffe erlaubt bereits ein recht gutes Resultat. Es besteht aber die Möglichkeit, dass mehr Bilder als bei den Unigrammen nicht bewertet werden können, weil deren Kommentare keines der sehr häufigen Wörter beinhalten könnten.

Vergleich mit manuell vergebenen Werten

Bisher wurden für die Schätzung die durch maschinelles Lernen generierten Schönheitswerte verwendet. Versuchsweise wurden für die 560 häufigsten Begriffe auch manuell

Schönheitswerte vergeben, um danach damit eine Schätzung durchzuführen. Das Resultat war mit einem r^2 von 0.356 (min. 50 Wörter und max. Stimmvarianz von 0.75) markant schlechter als mit den aus den Daten gewonnenen Werten. Das Vergeben von manuellen Ästhetikwerten ist also nicht trivial, obwohl aus der Behandlung der ersten Forschungsfrage bereits Vorwissen über die Charakteristik der Schönheitswerte besteht. Dies stützt die Vorgehensweise mit den datenbasierten Schönheitswerten.

6.2.6 Bigramme

Bei den bisherigen Schätzungen wurde der Satzstruktur, etwa Verneinungen oder Steigerungen von Adjektiven, in der Schätzung keine Beachtung geschenkt. Eine Möglichkeit, der Satzstruktur Rechnung zu tragen, ist die Verwendung von Bigrammen, also Wortpärchen, die mit jeweils zwei aufeinander folgenden Worten gebildet werden. So erhält das Bigramm *not bad* eine Bewertung von 8.25, während dem im Unigrammlexikon *not* nicht berücksichtigt wird, weil es ein sehr häufiges Wort ist und *bad* einen Durchschnitt von 5.175 erhält.

Um den Einfluss der verschiedenen Parameter auf die Bigramm-Schätzung zu evaluieren, wurden 100 Kombinationen der gewählten Parameter durchgerechnet. Wie schon bei der Kookkurrenz wurde für den Exponent m der Gewichtungsformel 0.5 eingesetzt, da dieser die besten Ergebnisse ermöglicht. Das Resultat dieser Berechnungen ist in der Abb. 6.20 dargestellt.

Das bereits bekannte Muster der abnehmenden Güte der Schätzung mit zunehmender maximaler Varianz der Stimmen ist auch hier wieder zu beobachten. Bei einer Mindestlänge des Kommentars von 20 Wörtern ist der Unterschied zwischen einer Varianz von 0.75 und 1.5 allerdings sehr klein. Auch der Trend des kleiner werdenden Bestimmtheitsmasses bei steigendem Mindestvorkommen im Trainingsset ist klar ersichtlich. Schliesslich verbessert sich die Schätzung bei länger werdenden Kommentaren, wie dies auch bei den Unigrammen der Fall ist.

Auffällig ist aber, dass bereits bei wenigstens 30 Worten im Kommentar ein r^2 von 0.482 erreicht wird und dieses bei mindestens 50 Worten nur auf 0.486 steigt. Die längeren Kommentare tragen folglich bei einem Mindestvorkommen im Trainingsset von einem Mal nicht zu einem besseren Resultat bei. Bei höheren Mindestvorkommen aber nimmt r^2 bei längeren Kommentaren wesentlich langsamer ab als bei kürzeren. Bei mindestens 25 Treffern im Trainingsset beträgt bei mindestens 50 Worten r^2 0.448, während es bei 30 Worten nur noch 0.397 erreicht (bei einer maximalen Varianz von 0.75).

Wenn die Auswahl der Bilder im Validierungsset nicht durch die maximale Varianz eingeschränkt wird, erreicht r^2 einen Höchstwert von 0.349 – im Vergleich zu den Unigrammen

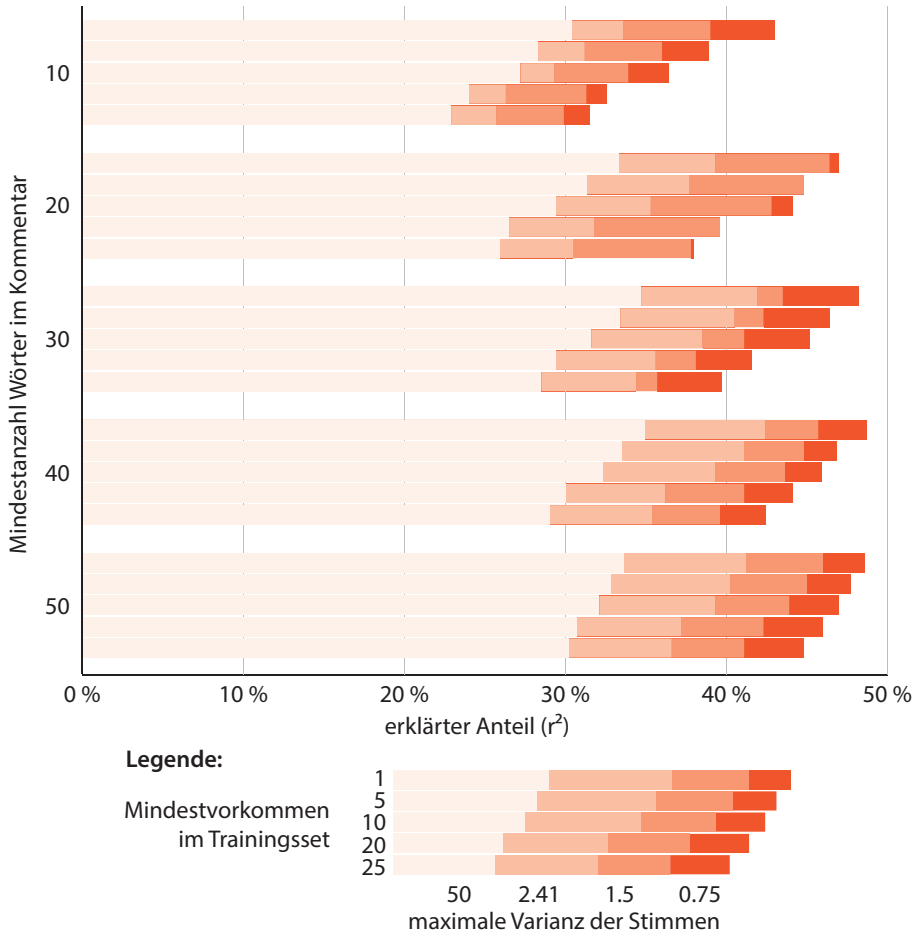


Abb. 6.20: Resultate der Bigramm-Schätzung. In horizontaler Richtung nimmt die maximal zugelassene Stimmvarianz ab (je dunkler desto kleiner); der Unterschied in vertikaler Richtung in den Fünfergruppen ist das Mindestvorkommen im Trainingsset.

ein etwas geringerer Wert, wo 0.385 erreicht wurde. Dieser Wert wird aber nicht bei den längsten Kommentaren, sondern bei einer Minimallänge von 40 Worten erreicht.

Vergleich mit Unigrammen und Kookkurrenz

Insgesamt kann das Bigrammschätzverfahren die Genauigkeit gegenüber der simpleren Unigrammmethode trotz der rudimentären Berücksichtigung der Satzstruktur nicht steigern, sondern bleibt etwas darunter. Es ist denkbar, dass der Korpus zur Bildung des Bigramm-Trainingssets zu klein ist, um den einzelnen Einträgen einen genügend repräsentativen Wert zu geben, denn zahlreiche Bigramme kommen sehr selten vor. Allerdings erreichen die Bigramme etwas bessere Resultate als das Kookkurrenzverfahren.

6.2.7 Trigramme

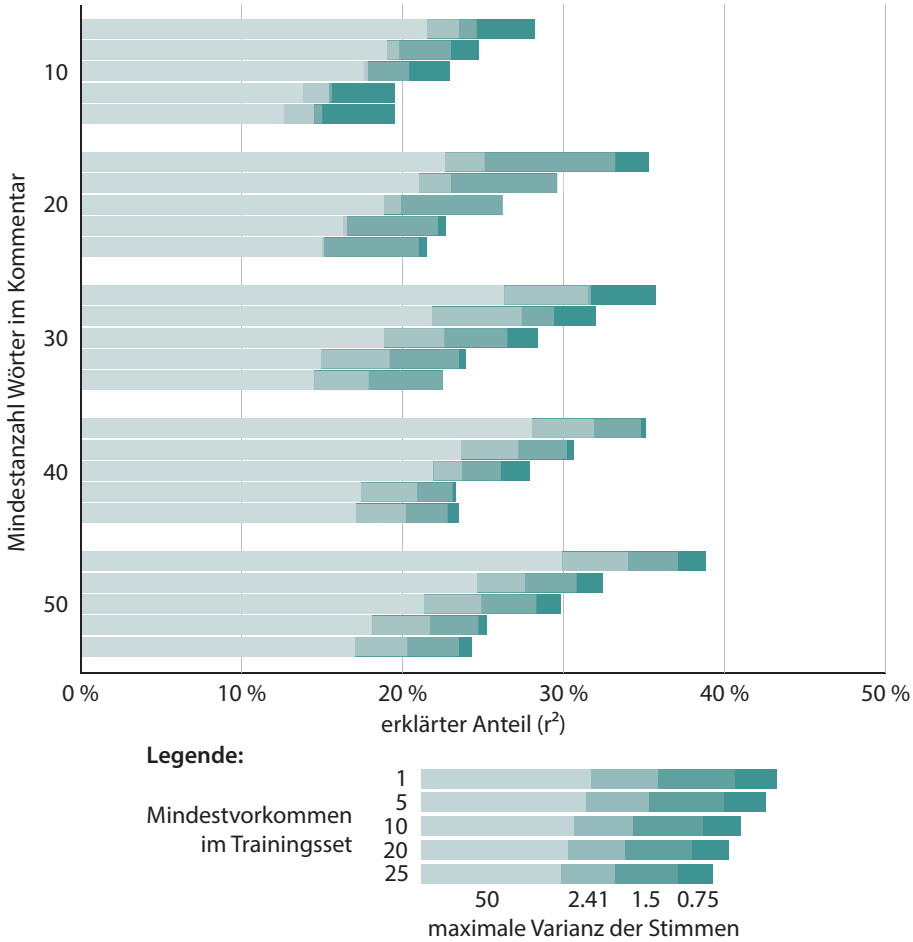


Abb. 6.21: Resultate der Trigramm-Schätzung. In horizontaler Richtung nimmt die maximal zugelassene Stimmvarianz ab (je dunkler desto kleiner); der Unterschied in vertikaler Richtung in den Fünfergruppen ist das Mindestvorkommen im Trainingsset.

Die Trigramme sind eine Fortsetzung der Bigramme, denn hier werden jeweils Dreiergruppen von aufeinanderfolgenden Worten betrachtet und zur Schätzung verwendet. Mit

dem Trigrammlexikon wurden dieselben Berechnungen wie mit dem Bigrammlexikon durchgeführt und dementsprechend ist auch die Grafik 6.21, die die Resultate der 100 durchgerechneten Kombinationen enthält, ähnlich aufgebaut.

Erneut sind die Trends des abnehmenden Bestimmtheitsmasses bei zunehmender maximaler Varianz der Stimmen und kürzer werdenden Kommentaren sichtbar. Teilweise sind die Unterschiede des Gütewertes bei den maximalen Varianzen 1.5 und 2.41 recht gering. Bei einer minimalen Kommentarlänge von 50 Worten tritt das Muster am deutlichsten auf.

Wenn die Anforderungen an das Mindestvorkommen im Trainingsset gesteigert werden, nimmt die Genauigkeit der Schätzung ab. Ob das Mindestvorkommen aber 20 oder 25 beträgt, spielt nur eine zu vernachlässigende Rolle, denn r^2 bleibt hier in etwa gleich. Dies hat damit zu tun, dass durch die Steigerung der Anforderung nicht viel weniger Trigramme im Trainingsset zur Verfügung stehen. Gesamthaft sind die Resultate der Trigrammschätzung eher enttäuschend, denn der beste erreichte Wert, 0.388 bei einer maximalen Varianz von 0.75, ist nur wenig über dem Wert, der bei den Unigrammen mit uneingeschränkter Varianz bei ansonsten gleichen Anforderungen erzielt wird (0.385). Auch gegenüber der Bigrammschätzung lässt sich eine deutliche Verschlechterung feststellen.

6.3 Flächendeckende Landschaftsevaluierung

Wie bereits erwähnt wurde, ist nur ein Bruchteil der Geograph-Bilder von ScenicOrNot bewertet worden. Die entwickelten Methoden können aber natürlich auch auf bisher nicht evaluierte Fotografien angewendet werden. So kann eine flächendeckende Ästhetikschätzung vorgenommen werden. Weil sich gezeigt hat, dass längere Kommentare bessere Ergebnisse erzielen, werden nur Bilder mit einer Textlänge von mindestens 200 Zeichen in diese Auswertung einbezogen. Die sogenannten ergänzenden Bilder von Geograph wurden ausgeschlossen, ebenso Bilder, die bereits von ScenicOrNot bewertet wurden. Damit stehen 280'092 Fotografien aus Grossbritannien zur Verfügung, die als Basis für die Textauswertung mit der Unigramm-Methode dienen. Es wurde keine geografische Auswahl getroffen, also beispielsweise die Anforderung, von möglichst allen Quadratkilometerzellen wenigstens ein Bild zu schätzen. Es gibt deshalb Rasterzellen, für die keine Schätzung vorgenommen wurde.

1.

http://www.geograph.org.uk/photo/2044450

Legende:
Porsche dealership located on Penarth Road.
The dealership is owned by Dick Lovett, who also has Ferrari and Maserati dealerships at the other end of the property.

1.

http://www.geograph.org.uk/photo/2040695

Legende:
Granite skerries off Inverliver Bay. View up the loch, with Beinn Trilleachan NN0843 on the left, and Ben Starav NN1242 on the right. In the distance are Bidean nam Bian NN1454 and Buachaille Etive Beag NN1753.

2.

http://www.geograph.org.uk/photo/748545

Legende:
Wickford Memorial Park.
Very impressive, take a look on Google Earth 51 37 05n, 0 32 16e. 80 acres of parkland,Cricket pitch,Football pitch,Six-rink bowling green Tennis Court,Basketball Court,Novelty golf,Children's playground, Sports pavilion,Riverside walk,Rose gardens, Small woodland and Memorial Avenue.

2.

http://www.geograph.org.uk/photo/940835

Legende:
Lamlash & the Arran hills taken whilst kayaking in Lamlash Bay. All 4 Corbetts on Arran can be seen: Beinn Tarsuinn-826m/2710ft, Cir Mhor-799m/2621ft, Caisteal Abhail -847m/2779ft & Goatfell-874m/2867ft.

3.

http://www.geograph.org.uk/photo/2385295

Legende:
Located at the southern end of Kingsway. Previously Nationwide Autocentre, the premises were rebranded after the purchase by Halfords of all 224 Nationwide Autocentres in February 2010.

3.

http://www.geograph.org.uk/photo/122232

Legende:
An attempt at the skyline left to right. Ben Starav (distant), Buachaille Etive Mor (triple pointy summit), Buachaille Etive Beag, Bidean Nam Bian, Aonach Eagach, Beinn a' Bhiethir (distant) and rightmost Sgurr Eilde Mor

Abb. 6.22: Geograph-Bilder mit den durch das Unigramm-Modell als hässlichsten (links) resp. schönsten (rechts) geschätzten Landschaften

In der Abb. 6.22 sind bisher nicht evaluierte Bilder dargestellt, die aufgrund des Textes als schönste bzw. hässlichste Landschaften geschätzt wurden. Es ist klar ersichtlich, dass die Schätzung plausibel ist. Weil die Schätzung für ganz Grossbritannien durchgeführt wurde, kann das Resultat mit den Bewertungen von ScenicOrNot verglichen werden. Damit der Vergleich gemacht werden kann, wurden die geschätzten Werte mit einer linearen Transformation auf den Bereich 1 bis 10 skaliert. Der Vergleich ist in der Abb. 6.23 dargestellt. Zur Darstellung wurden die einzelnen Fotostandpunkte zu 5 * 5 Kilometer grossen Rasterzellen zusammengefasst. Für jede Rasterzelle wurde der Durchschnitt der Bewertungen berechnet. Auf der linken Seite sind die Bewertungen von ScenicOrNot zu sehen, auf der rechten Seite die nur aufgrund der Unigramm-Textanalyse erstellten Schätzungen.

Man erkennt, dass das Modell die generelle Verteilung der Bewertungen gut nachbilden kann. Es ist also möglich, das räumliche Muster der Landschaftsschönheit nur aufgrund von Worten zu modellieren. Bei den ScenicOrNot-Bewertungen hat fast jede Rasterzelle einen Wert, bei der Modellierung fehlen aber für etliche Zellen die Werte. Der Grund dafür ist, dass bei der Auswahl der Bilder die Länge der Kommentare als Kriterium festgelegt wurde und nicht eine möglichst vollständige Abdeckung Grossbritanniens angestrebt wurde.

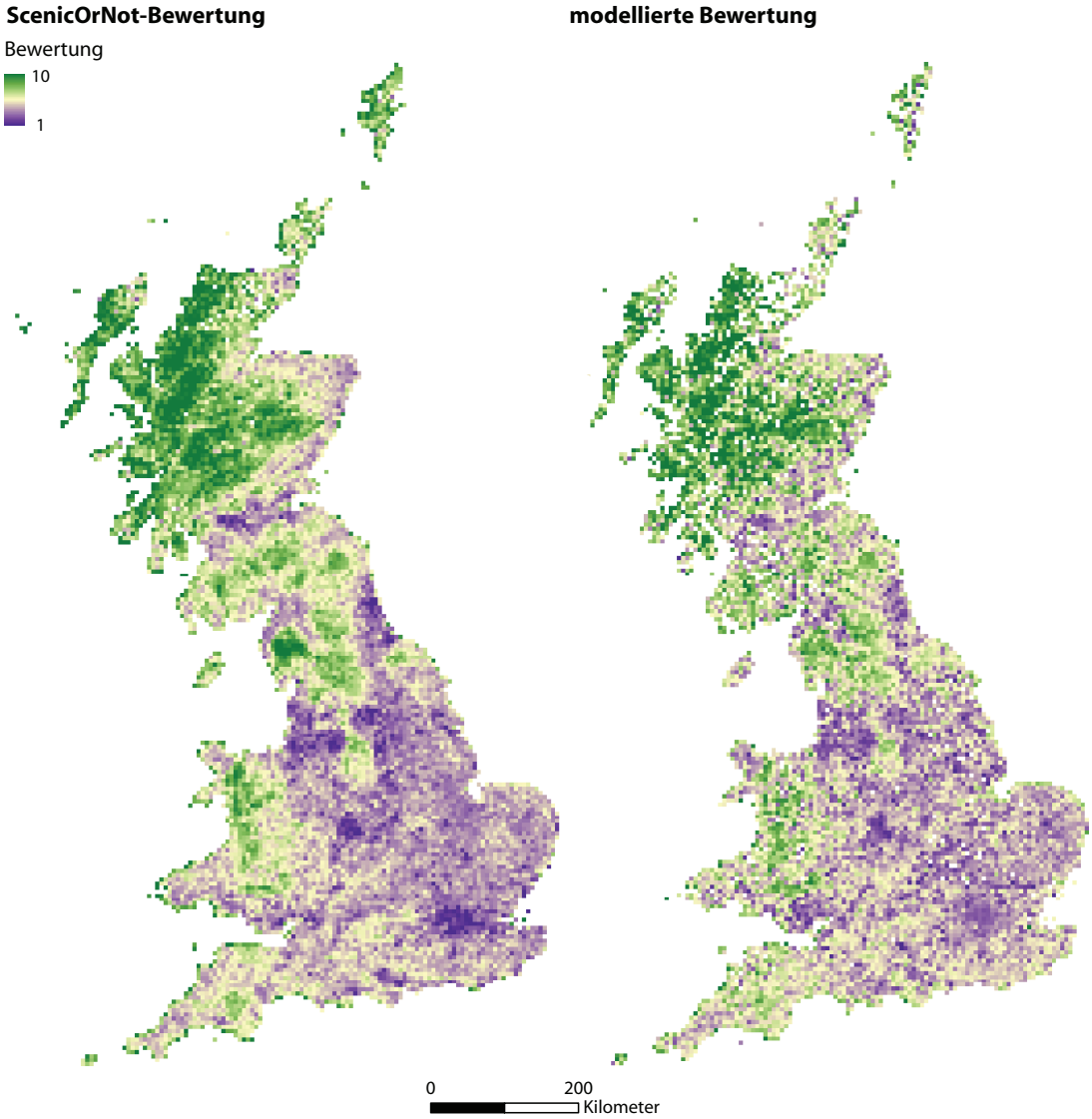


Abb. 6.23: Vergleich der ScenicOrNot-Bewertung (links) und der modellierten Bewertung aufgrund der Texte von 280'000 Bildern (rechts) in einer 5km-Auflösung. Man erkennt, dass das generelle Muster durch das Modell gut wiedergegeben wird.

7 Diskussion

In diesem Kapitel werden die angewendeten Vorgehensweisen und die erzielten Ergebnisse kritisch betrachtet. Zuerst werden einige Anmerkungen zu den Methoden zur Erkennung der Muster in den Daten gemacht, anschliessend werden die Resultate der Bewertungsschätzung erörtert.

7.1 Muster in den Daten

Die Behandlung der ersten Forschungsfrage basiert stark auf einer explorativen Datenanalyse, die es ermöglicht hat, die versteckten Muster der verwendeten Datensätze sichtbar zu machen. Anfangs bestand die Vermutung, zur Beantwortung der zweiten Forschungsfrage sei ein frei verfügbarer, generischer Stimmungsdatensatz wie SentiWordNet geeignet. Erst die während der explorativen Analyse identifizierten Muster wiesen auf die Untauglichkeit des SentiWordNet-Datensatzes hin, weil auch Wörter, die in Bezug auf die menschliche Stimmung neutral sind, eine Information über die Landschaftsschönheit enthalten. Aufgrund dieser Erkenntnis wurde ein spezifisch für die gestellte Aufgabe geeignetes Schönheitslexikon aus den Daten generiert.

Neben den Erkenntnissen über die Struktur der Geograph-Texte sind die gewonnenen Einsichten vor allem für die Planung des weiteren Vorgehens zur Beantwortung der zweiten Forschungsfrage sehr hilfreich.

7.1.1 Wortwolken

Wortwolken sind eine Visualisierungsmethode, bei der häufig vorkommende Begriffe in einem Datensatz grösser als seltenere Worte dargestellt werden. Die Wortwolken basieren auf willkürlich gesetzten Grenzwerten, die die Begriffe in eine ebenfalls willkürlich gewählte Anzahl verschiedene Kategorien aufteilen. Die gewählten Werte sind aber sinnvoll und erlauben eine gute Übersicht über die Charakteristik der Geograph-Daten. Problematisch ist, dass trotz der grossen Anzahl Bilder, die ausgewertet wurden, einige Begriffe sehr selten gefunden wurden. Dies ist insbesondere in den hohen Bewertungskategorien und noch stärker bei den hoch bewerteten Adjektiven der Fall. Ausserdem dominieren Begriffe mit extrem häufigem Auftreten die Darstellung, was zur Folge hat, dass seltene Begriffe beinahe verschwinden und kaum mehr erkennbar sind.

Das Tool zur Erstellung der Wortwolken (www.wordle.net) wurde von *Viegas et al.* (2009) beschrieben. Diese Autoren haben auch eine Umfrage über dieses Werkzeug durchgeführt. Gemäss dieser Umfrage hat die ausgefeilte Typografie den Effekt, dass die Betrachter die Kreationen länger und genauer anschauen. Dies ist ein wünschenswerter Effekt. *McNaught und Lam* (2010) haben ebenfalls Wordle benutzt, um Antworten, die im Rahmen von zwei Forschungsprojekten entstanden sind, zu visualisieren. Ihre Schlussfolgerungen sind, dass Wordles gut geeignet sind, um sich einen ersten Überblick über textuelle Daten zu verschaffen. Dabei sollten die Wordles aber nur als Zusatzwerkzeug und nicht als *stand-alone-tool* eingesetzt werden. Dies wurde auch in dieser Masterarbeit so gehandhabt.

Die Attribute, die visuell ansprechenden bzw. nicht ansprechenden Landschaften zugeschrieben werden, wurden von *Craik* (1972b) untersucht. Die bei der Wordle-Erstellung berechneten Werte decken sich mit seinen Resultaten, denn die Attribute ansprechender Landschaften erzielen im Durchschnitt höhere Werte als die Attribute nicht ansprechender Landschaften.

7.1.2 Multidimensionale Skalierung

Die multidimensionale Skalierung ist ein geeignetes Verfahren, um die verwendeten Daten zu visualisieren. Weil es unmöglich ist, alle in den Geograph-Kommentaren gefundenen Begriffe im Ergebnis einer multidimensionalen Skalierung darzustellen, muss eine Auswahl der auszuwertenden Begriffe getroffen werden. Für diese Auswahl gibt es keine richtige oder falsche Lösung, aber es bietet sich an, Begriffe mit eher hohem Vorkommen im Datensatz zu wählen, weil die multidimensionale Skalierung auf dem gleichzeitigen Auftreten derselben fusst. Es wären aber fast unbeschränkt viele Kombinationen möglich und die Auswahl ist wie bei den Wortwolken willkürlich.

Visualisierungen mit multidimensionalen Skalierungen wurden auch von anderen Autoren benutzt, um sich einen Überblick über die in einem Textkorpus behandelten Themen zu verschaffen, z. B. von *Fortuna et al.* (2005), die damit thematisch ähnliche Wörter erforscht haben. Wie in der Abbildung 6.9 der multidimensionalen Skalierung zu sehen ist, liegen Begriffe mit ähnlicher Thematik in der Nähe voneinander. *Lund und Burgess* (1996) haben die auch in dieser Arbeit verwendete Methode des gleichzeitigen Auftretens von Wortpaaren verwendet, um mit multidimensionaler Skalierung eine Wortkarte zu erstellen. *Wise* (1999) hat zur Visualisierung seines Textkorpus die auch im vorliegenden Werk eingesetzte Landschaftsmetapher verwendet, allerdings mit einer dreidimensionalen Ansicht. In dieser Masterarbeit wurde bewusst auf eine dreidimensionale Darstellung verzichtet, um die Lesbarkeit zu erhöhen. Stattdessen wurde zur Andeutung der Topografie, die die Schönheitswerte wiedergibt, auf Höhenlinien gesetzt.

7.2 Bewertungsschätzung

Die Bewertungsschätzungen beruhen auf der Analyse der Begleittexte mit verschiedenen Methoden und dem anschliessenden Vergleich des geschätzten Wertes mit dem wahren Wert. Nachfolgend werden die Beobachtungen, die sich aus den Resultaten ergeben, erörtert.

7.2.1 Maximale Stimmvarianz

Bei allen Methoden zur Landschaftsbewertungsschätzung hat sich gezeigt, dass sich der erklärte Anteil verbessert, wenn die Auswahl des Validierungssets durch die maximal zugelassene Stimmenvarianz in den ScenicOrNot-Daten eingeschränkt wird. Durch diese Einschränkung werden einerseits Bilder herausgefiltert, bei deren Bewertung die ScenicOrNot-Skala teilweise falsch interpretiert wurde, denn diese haben per Definition eine hohe Varianz und machen einen Vergleich mit einem geschätzten Wert unsinnig. Gleichzeitig ist die maximale Stimmvarianz aber auch ein Mass dafür, wie stark sich die Bewertenden einig waren. Offensichtlich ist Schätzung von Bildern, bei deren Bewertung Einigkeit geherrscht hat, präziser.

Die Begrenzung der maximalen Stimmvarianz ist eine relativ starke Einschränkung, denn dieses Wissen besteht, im Gegensatz zur Kommentarlänge, nur weil die Bilder bereits bewertet wurden. Wenn die Schätzung auf nicht bewertete Bilder, etwa von Flickr, ausgeweitet würde, ist diese Information nicht bekannt.

Die unterschiedlichen Präferenzen in der Landschaftsbeurteilung wurden beispielsweise von *Van den Berg et al.* (1998) erforscht. Dabei hat sich gezeigt, dass sich die Bewertungen zwischen Landwirten, Einheimischen und Touristen unterscheiden. Die Forscher vermuten, dass die Unterschiede aufgrund von unterschiedlichem sozialen Hintergrund und der Vertrautheit mit der Landschaft zustande kommen. Bei den ScenicOrNot-Daten sind die Gründe für unterschiedliche Bewertungen unbekannt, weil man über die Bewerter nichts weiss.

Ein Teil der festgestellten Stimmvarianz in den ScenicOrNot-Daten könnte auch von der Fotoqualität herrühren und wenig mit der Landschaftsästhetik zu tun haben, denn die Geograph-Bilder wurden nicht unter kontrollierten Bedingungen geschossen und können sich in ihrer Qualität unterscheiden.

7.2.2 Transformation der Werte

Zur Berechnung des geschätzten Schönheitswertes wurde nicht der ursprüngliche Durchschnittswert eines Begriffes verwendet, sondern eine Transformation desselben. Durch die Transformation, die den Kontrast zwischen guten und schlechten Bewertungen erhöht, verbessert sich das Bestimmtheitsmass der Schätzung. Der Grund für die Verbesserung ist schwierig zu eruieren. Offenbar ist es hilfreich, wenn die Werte der einzelnen Begriffe nicht zu nahe am Mittelwert liegen, sondern zu den Extremen hin transformiert werden.

7.2.3 Toponymerkennung

Sowohl bei der Erstellung der Lexika als auch bei der Schätzung können durch die Toponymfilterung Fehler entstehen. Einerseits werden Toponyme nicht als solche erkannt, wenn sie beispielsweise zu weit vom Fotostandort entfernt oder nicht im Gazetteer erfasst sind. Bei der Lexikonerstellung erhalten dadurch zu Toponymen gehörende Wörter ebenfalls Stimmungswerte, die später nicht herausgefiltert werden können. Bei der Schätzung wiederum werden diese nicht detektierten Toponyme zur Schätzung verwendet. Umgekehrt können auch Wörter als Toponym klassiert werden, obwohl sie es eigentlich nicht sind. Es hat sich gezeigt, dass der Einbezug von Toponymen die Schätzung verbessert – Fehler in der Toponymerkennung haben deshalb keinen negativen Einfluss auf die Qualität der Schätzung. Toponyme sind aber ein starker Hinweis auf die Schönheit eines Ortes: Es entstehen sehr unterschiedliche Bilder im Geist, wenn man beispielsweise «Jungfraujoch» oder «Schlieren» hört. Ohne Toponymfilterung bei der Lexikonerstellung erhalten solche Ortsnamen ebenfalls Schönheitswerte. Das Interesse, die Toponyme herauszufiltern besteht deshalb darin, auch ohne die Hilfe der Ortsnamen eine möglichst gute Schätzung der Schönheit eines Ortes zu machen.

In den kurzen Texten der Geograph-Kommentare kommen typischerweise wenige Toponyme vor – meistens zwischen null und vier. Dementsprechend schwankt die Trefferquote zwischen 0 % wenn etwa das einzige vorkommende Toponym nicht erkannt wurde und 100 %. Wie sich gezeigt hat, ist vor allem die Erkennung der schottischen Toponyme, insbesondere die Namen der Berge, anspruchsvoll. In der Abb. 7.1 ist die Dichte der Toponyme in Rasterzellen mit 5 km Seitenlänge dargestellt. Die Werte pro Zelle variieren zwischen einem und 66 Toponymen. In Schottland ist die Dichte geringer als in den meisten anderen Regionen des Untersuchungsgebietes.

Viele Autoren befassen sich mit den Mehrdeutigkeiten von Toponymen, also dass verschiedene Orte denselben Namen tragen, so beispielsweise *Gelbukh et al.* (2004) und *Samet* (2011). In der vorliegenden Arbeit ist dies weniger von Interesse, denn einerseits sind die Koordinaten der Bilder vorhanden und vor allem geht es nicht darum, auf-

grund der im Text gefundenen Toponyme den Ort zu referenzieren. Problematisch sind eher Mehrdeutigkeiten die entstehen, wenn ein Wort wie *hill* sowohl Ortsname als auch Landschaftsmerkmal sein kann.

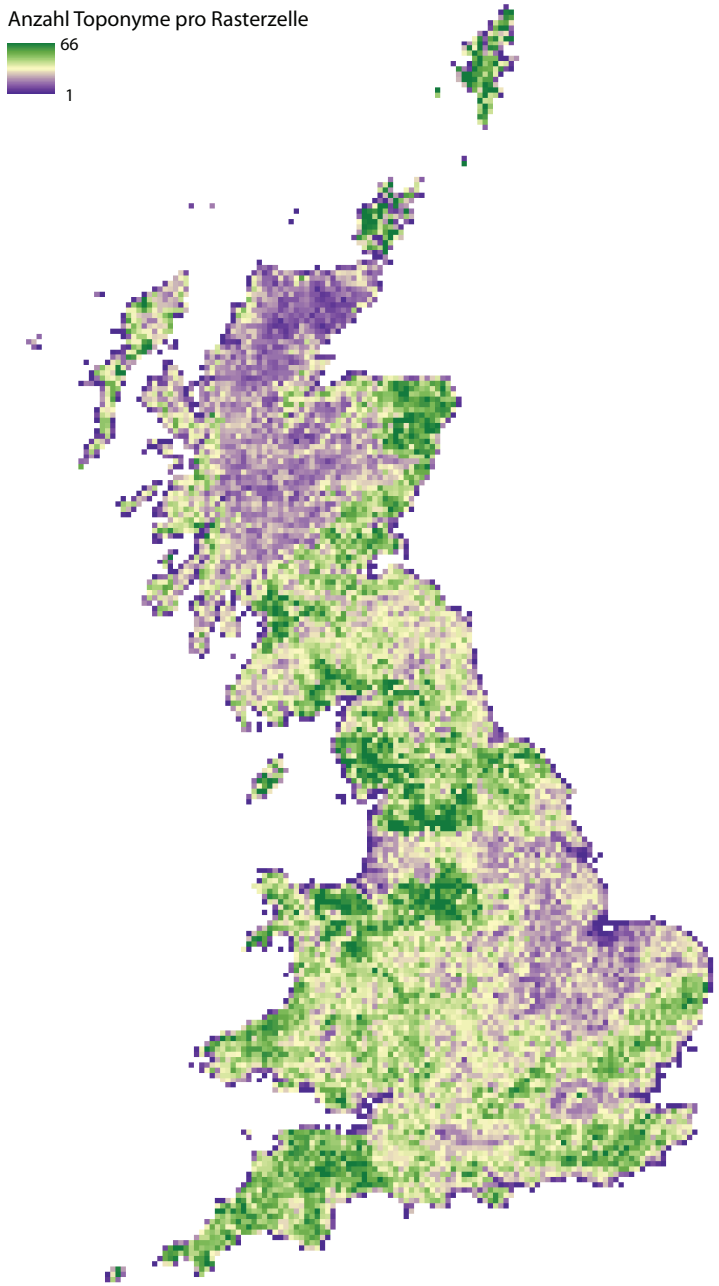


Abb. 7.1: Toponymdichte in Grossbritannien in 5km-Rasterzellen. Quelle: Gazetteer des Ordnance Survey GB, eigene Darstellung

7.2.4 Unigramme

Erstaunlicherweise wurden mit den Unigrammen die besten Resultate erzielt. Dies ist deswegen bemerkenswert, weil mit dieser Methode die Satzstruktur komplett ignoriert wird – Verneinungen und dergleichen werden also nicht erkannt. Der Vorteil der Unigramm-Methode besteht darin, dass das benutzte Stimmungslexikon sehr viele Begriffe enthält. Wie sich gezeigt hat, verbessert sich die Schätzung, wenn mehr Begriffe dafür zur Verfügung stehen. Damit wurde die Vermutung, dass sich das Resultat bei höherer Anforderung an das Vorkommen im Lexikon verbessert, widerlegt, weil dann die Anzahl der zur Verfügung stehenden Begriffe abnimmt. Dasselbe ist auch bei der Anwendung des Unigrammlexikons, das nur auf Bildern mit geringen Varianzen basiert, zu beobachten: Es bestand die Hypothese, dass die höhere Sicherheit in der Bewertung der einzelnen Begriffe ein höheres Bestimmtheitsmass der Schätzung erlauben würde. Die Experimente haben aber gezeigt, dass die Verkleinerung des Lexikons generell zu einer Verschlechterung des Resultats führt.

Schliesslich ist dieser Zusammenhang auch bei der Schätzung mit den 560 häufigsten Begriffen erkennbar: Im Vergleich zur vollständigen Unigrammschätzung ist der erklärte Anteil etwas geringer. Allerdings erlauben die 560 Begriffe bereits eine erstaunlich präzise Schätzung. Ausserdem ermöglicht es der begrenzte Umfang der sehr häufigen Begriffe, die Schönheitswerte auch manuell zu vergeben. Allerdings konnte kein Vorteil der manuellen Wertvergabe gegenüber dem maschinellen Lernen festgestellt werden.

Es wurde auch versucht, die Stimmungswerte innerhalb des Untersuchungsgebiets geografisch differenziert anzuwenden. Damit konnte keine Verbesserung gegenüber den normalen Unigrammen erreicht werden. Es ist unklar, ob sich die Resultate nur aufgrund des zu kleinen Lexikons verschlechtern oder ob die Annahme, dass sich die Stimmungswerte auf kleinräumigen Gebiet überhaupt unterscheiden, falsch ist.

Vergleich mit der Literatur

Stadler (2010) konnte 51.8 % der Varianz in den Schönheitsbewertungen von ScenicOrNot mit den Bodenbedeckungsklassen erklären und diesen Wert durch Ausschluss von Klassen mit wenig Erklärungsgehalt auf 62.6 % steigern. Bei Einschränkung der maximalen Stimmvarianz konnte mit der Unigramm-Methode ebenfalls mehr als die Hälfte der Variation von ScenicOrNot erklärt werden, bei uneingeschränkter Varianz immerhin 39.1 %. *Shafer und Mietz* (1970) haben ein Modell zur Schätzung der ästhetischen Qualität einer Landschaft erstellt, das auf der Analyse der auf einer Fotografie eingenommenen Fläche eines Landschaftselements basiert. Sie berichten, dass sie 66 % der Variation mit ihrem Modell erklären konnten. Dies ist deutlich mehr als in der vorgestellten Textanalysierungsmethode,

aber die Flächenabmessungen wurden manuell vorgenommen. Auch heutzutage dürfte die automatische, computerbasierte Erkennung von Landschaftselementen auf Bildern sehr schwierig sein. Der Vorteil der Textanalyse besteht darin, dass auf einfache Weise umfangreiche Datensätze ausgewertet werden können.

7.2.5 Bi- und Trigramme

Wie sich bei den Resultaten gezeigt hat, werden bei den Bi- und Trigrammschätzungen schlechtere Werte als bei der Unigrammschätzung erreicht. Es ist denkbar, dass der Grund dafür in den verhältnismässig kleinen Bi- und Trigrammlexika liegt. Das Bigrammlexikon umfasst circa 300'000 Wortpärchen und das Trigrammlexikon enthält rund 650'000 Wortgruppen. Diese Zahlen kann man in einen Bezug zu einem theoretischen Total setzen, obwohl es schwierig ist, eine Gesamtzahl der Wörter in der englischen Sprache anzugeben, weil es keine Regeln gibt, wie die Wörter genau zu zählen sind. In der Literatur findet man Werte zwischen 88'500 (*Nagy und Anderson* 1984) und über einer Million Wörter (*Michel et al.* 2011). Bei der konservativen Schätzung wären also theoretisch fast 8 Milliarden (88500^2) Bigramme möglich. Sicherlich machen bei weitem nicht alle dieser möglichen Kombinationen Sinn, trotzdem zeigt es, wie klein die beiden Lexika sind. Es ist deshalb sehr gut möglich, dass etliche Wortgruppen, die im Validierungsset gefunden wurden, kein Pendant im Trainingsset haben, weil dieses zu wenig umfangreich ist. Bei den Unigrammen wurde gezeigt, dass ein möglichst umfassendes Lexikon die besten Resultate erzielt und es wäre deshalb auch für die Bi- und Trigrammschätzung von Vorteil, ein umfassenderes Lexikon zur Verfügung zu haben. Im Abschnitt 7.4.4 ist ein Vorschlag zur Vergrösserung der Lexika zu finden.

Bei der Unigrammschätzung wurden sehr häufige Wörter in den Kommentaren wie *the* bei der Schätzung nicht berücksichtigt. Bei den Bi- und Trigrammen ist dies weniger einfach, denn es ist unklar, wie extrem geläufige Wortgruppen zu definieren sind.

Ein Vorteil der Bigrammschätzung besteht darin, dass bereits bei kürzeren Kommentaren (ab 30 Worten) ein ähnlich guter Wert wie bei längeren Texten erreicht wird. Dies ist insbesondere bei der Unigrammschätzung nicht der Fall. Die Bigrammschätzung scheint also bei kürzeren Texten besser zu funktionieren.

Dave et al. (2003) haben in einer Opinion-Mining-Studie unter anderem Uni- und Bigramme verglichen. Die Bigramme haben dabei jeweils leicht bessere Werte erzielt. In der vorliegenden Studie konnte dieses Ergebnis nicht reproduziert werden.

7.2.6 Kookkurrenz

Im Gegensatz zu den n-gramm-Lexika, deren Bestwerte jeweils bei keinen Einschränkungen bezüglich des minimalen Vorkommens im Trainingsset zustande kamen, erreicht die Kookkurrenz-Schätzung die besten Resultate, wenn die Begriffspaare mindestens zwischen 5 und 10 mal vorkommen. Offenbar ist die Bewertung der sehr selten gefundenen Begriffspaare unsicher und der Einbezug dieser verschlechtert die Schätzung.

Das Kookurrenzlexikon beruht auf den häufigsten Begriffen des Datensatzes, was den Vorteil hat, dass viele Begriffspaare gefunden werden. Es gibt aber keine richtige Lösung, welche Begriffe ins Kookkurrenzlexikon einbezogen werden sollen und die häufigsten sind nicht zwingend auch die aussagekräftigsten.

Weil die Resultate der Kookkurrenzschätzung sowohl der Uni- wie auch der Bigrammschätzung unterlegen sind und ausserdem höhere Anforderungen an das Trainingsset stellen (Mindestvorkommen), erscheint eine weitere Verfolgung dieses Verfahrens bei Schätzung mit ganzen Sätzen als Datengrundlage nicht sinnvoll.

Turney (2002) hat ein kookkurrenzbasiertes Verfahren zur Einstufung von Reviews verwendet. Er hat damit eine durchschnittliche Genauigkeit von 74 % erzielt, aber seine Einstufung basiert auf einer binären Skala (empfohlen / nicht empfohlen).

7.3 Einschränkungen

7.3.1 Abhängigkeit vom Untersuchungsgebiet

Das Untersuchungsgebiet dieser Arbeit ist Grossbritannien. Weil das Schönheitslexikon auf diesem Gebiet beruht, dürfte sich die Übertragung auf andere Gebiete nicht problemlos gestalten, denn im Lexikon sind die regionalen Eigenheiten abgebildet. So sind viele der positiven Begriffe schottische Wörter, die andernorts kaum vorkommen dürften. Vor einer Anwendung auf eine andere Region müsste das Bewertungslexikon deshalb mit dortigen lokalen Begriffen ergänzt werden. Ausserdem ist das Bewertungslexikon auch von der Geomorphologie abhängig. Da zum Beispiel Wüsten in Grossbritannien fehlen, hat das in Geograph zwar vorkommende Wort *oasis* wohl eine andere Bedeutung und damit auch einen anderen Schönheitswert als wenn es beispielsweise in Australien oder Südafrika verwendet wird.

Gleichzeitig ist anzumerken, dass die manuelle Vergabe von Schönheitswerten nicht trivial ist und der Bestimmung durch Schwarmintelligenz tendenziell unterlegen ist, wie sich bei Experimenten mit den häufigsten vorkommenden Begriffen gezeigt hat.

7.3.2 Ganze Sätze

Die Bewertungsschätzung wurde in dieser Arbeit auf der Basis von ganzen Sätzen durchgeführt. Viele Datensätze im Internet, deren Auswertung geprüft werden könnte, operieren aber lediglich mit Stichworten (*tags*). Es ist unklar, welche Resultate durch Anwendung der Methodik auf einen Datensatz mit Stichworten anstelle von ganzen Sätzen erzeugt werden könnten. Es wäre aber interessant, dies in einer weiterführenden Arbeit zu prüfen.

7.3.3 Berücksichtigung der Polyanna-Hypothese

Gemäss der Polyanna-Hypothese (*Boucher und Osgood* 1969) ist die menschliche Sprache positiv verzerrt. Einige Autoren gewichten deshalb die seltener vorkommenden negativen Begriffe stärker. Die hier verwendeten Stimmungswerte beziehen sich aber nicht auf die menschliche Stimmung und es ist deshalb nicht evident, dass negativ bewertete Landschaftsbegriffe stärker gewichtet werden sollten. Abgesehen davon würde sich die Abgrenzung, ab welcher Bewertung Begriffe stärker gewichtet werden sollten, schwierig gestalten. Aus diesen Gründen wurde auf eine solche Berechnung verzichtet.

7.4 Mögliche Verbesserungen

7.4.1 Rechtschreibkorrektur

Soweit dies auf der Geograph-Website ersichtlich ist, haben die Geograph-Kommentare keine redaktionelle Betreuung. Schreibfehler oder auch fehlende Leerschläge können deshalb auftreten. Weder bei der Lexikonbildung für das Trainingsset noch bei der Schätzung der Bewertungen wurde dies berücksichtigt. Es wäre deshalb möglich, sowohl bei der Bildung des Lexikons als auch während der Schätzung eine automatische Rechtschreibkorrektur vorzunehmen. Eine Implementierung könnte bei Worten, die nicht in einem Thesaurus (beispielsweise der in dieser Arbeit verwendete *Moby Thesaurus*) gefunden werden, mithilfe der Levenshtein-Distanz das nächste bekannte Wort bestimmen und dann dieses verwenden. Für fehlende Leerschläge müsste eine andere Lösung gesucht werden. Beispielsweise könnte die urspüngliche Zeichenkette zwischen jedem Zeichen einmal aufgespalten werden und die für die so entstehenden Wörter würde dann jeweils geprüft, ob sie im Thesaurus vorkommen. Dies ist aber ein zeitaufwendiges Vorgehen, weil es nach dem Trial-and-Error-Prinzip vorgeht. Gemäss *Hodge und Austin* (2002) ist auch der Levenshtein-Algorithmus rechenintensiv, weil er einen Brute-force-Ansatz verfolgt. Sie schlagen deshalb in ihrem Paper eine andere Vorgehensweise vor. *Kukich* (1992) gibt einen umfangreichen Überblick über verschiedene Verfahren zur automatischen Rechtschreibkorrektur.

7.4.2 Berücksichtigung der Satzstruktur

In dieser Arbeit wurde versucht, die Satzstruktur mithilfe von Bi- und Trigrammen abzubilden, allerdings nur mit mässigem Erfolg. Ein anderer Ansatz wäre die Verwendung eines Unigramm-Lexikons und zusätzlich eine regelbasierte Erkennung der Satzstruktur. Wenn also etwa vor einem Adjektiv ein *not* entdeckt wird, müsste der gegenteilige Wert des Adjektivs verwendet werden. Mit einer Reihe solcher Regeln könnte das Resultat möglicherweise verbessert werden.

Lambek (1958) hat sich bereits in den 1950er-Jahren der mathematischen Analyse von Satzstrukturen gewidmet. Ein Beispiel für die Anwendung von Satzstrukturerkennung ist in der Arbeit von *Athar* (2011) zu finden. Dieser Autor hat Sentiment Analysis auf wissenschaftliche Publikationen angewendet, um herauszufinden, ob Forschungsresultate bei der Zitierung Unterstützung erfahren oder ob ihnen widersprochen wird. Der Vorteil der Strukturerkennungsverfahren ist, dass sie auch weit auseinanderliegende Bezüge zwischen Worten, die Trigramme nicht entdecken, erkennen können.

7.4.3 Erkennung des Subjekts

In dieser Arbeit wurde der gesamte Kommentar eines Bildes zur Schätzung einbezogen. Gerade bei längeren Kommentaren kann es vorkommen, dass Details zur Geschichte eines abgebildeten Ortes oder andere irrelevante Dinge wiedergegeben werden, die bei der Bestimmung der Ästhetik nicht hilfreich sind. Mit Methoden des *natural language processing* könnte versucht werden, satzweise zu bestimmen, welches Thema im Text angesprochen wird, und danach nur relevante Sätze zu berücksichtigen. Diese Aufgabe ist mit der Erkennung von Geräteeigenschaften beim Opinion Mining einzelner Aspekte verwandt, wie sie in der Literatur bei Reviewanalysen (z. B. *Yi et al.* 2003) beschrieben wird.

7.4.4 Erweiterung des Bi- und Trigrammlexikons

Bei der Diskussion der Bi- und Trigrammschätzung wurde erwähnt, dass diese Lexika möglicherweise zu klein sind. Eine Möglichkeit zur Vergrösserung dieser Lexika besteht darin, die Ästhetik weiterer, noch nicht bewerteter Geograph-Bilder mit dem Unigrammverfahren zu schätzen und mit den so evaluierten Bildern aus den Kommentaren neue Wortgruppen zu generieren und Schönheitswerte zu vergeben. Das so entstehende Lexikon hätte aber den Nachteil, dass es nur auf geschätzten Landschaftsbewertungen basieren würde. Die automatische Erweiterung von Lexika wird auch in der Literatur beschrieben, so etwa von *Perrie et al.* (2013). Diese Autoren verwenden aber als Basis von Menschen

vergebene Werte, während in der vorliegenden Arbeit nur die zugrunde liegenden Landschaftsbewertungen von Menschen erstellt wurden. Die Schönheitswerte aber sind bereits durch eine unüberwachte Klassierung generiert worden.

7.5 Beantwortung der Forschungsfragen

Forschungsfrage 1: Welche Muster kann man in den Kommentaren der Geograph-Bilder, die von ScenicOrNot bewertet wurden, entdecken?

Die Auswertung der Bildkommentare mit den ScenicOrNot-Bewertungen zeigt, dass Begriffe mit urbaner Konnotation deutlich schlechtere Schönheitsbeurteilungen erhalten als Begriffe des Naturraums. Es hat sich gezeigt, dass zahlreiche zur Bewertung von Landschaften wichtige Begriffe in einem generischen Stimmungslexikon als neutral eingestuft werden, insbesondere auch Nomen. Ausserdem werden viele Begriffe in einem landschaftlichen Kontext auf sehr spezifische Art und Weise eingesetzt, so zum Beispiel *gash* (Schnittwunde), das in Geograph für positiv bewertete Landschaftseinschnitte verwendet wird. Aufgrund dieser Tatsachen wurde klar, dass ein generisches Stimmungslexikon zur Beantwortung der zweiten Forschungsfrage ungeeignet ist. Es wurden deshalb auf der Basis eines Trainingssets landschaftsspezifische Lexika erstellt, deren Leistungsfähigkeit in der zweiten Forschungsfrage überprüft wurde.

Forschungsfrage 2: Welcher Anteil der Varianz einer Landschaftsbewertung kann durch eine Analyse einer Bildlegende oder eines Begleittextes erklärt werden?

Die Schätzungen der Landschaftsbewertungen basieren auf den Lexika mit Schönheitswerten für einzelne Worte bzw. Wortkombinationen. Zur Schätzung der Landschaftsbewertungen wurden verschiedene Methoden entwickelt, deren Resultate im Folgenden konzis dargelegt werden. Generell verbessert sich die Schätzung, wenn die maximal zulässige Stimmenvarianz der Bilder eingeschränkt wird und bei den meisten Methoden wird das Resultat umso besser, je länger der Begleittext ist (bis zur untersuchten Gesamtlänge von 50 Worten). Diese Einschränkung kann natürlich nur bei der Validierung vorgenommen werden, wenn der wahre Wert bekannt ist. Bei der Schätzung bisher nicht evaluierter Bilder gilt für die Bestimmtheitsmasse deshalb der Wert für uneingeschränkte Stimmvarianz.

	max. Stimmvarianz 0.75	max. Stimmvarianz unbeschränkt
Unigramme	0.532	0.391
Kookkurrenz	0.455	0.328
Bigramme	0.487	0.349
Trigramme	0.388	0.299

Tabelle 7.1: Übersicht über die erreichten Bestimmtheitsmasse r^2 mit den verschiedenen Methoden

Wie in der Tabelle 7.1 ersichtlich ist, konnte mit der Auswertung von Bildlegenden durch das Unigramm-Modell maximal etwas mehr als die Hälfte der Varianz der Bewertungen erklärt werden, wenn die maximale Stimmvarianz eingeschränkt wird. Bei unbeschränkter Stimmvarianz sinkt der maximal erklärte Anteil auf 39.1 %. Die Bigrammmethode kann jeweils ca. 5 Prozentpunkte weniger erklären und das Trigrammverfahren schneidet am schlechtesten ab.

8 Schlussfolgerungen

8.1 Was wurde erreicht?

Das Ziel dieser Arbeit war, zu erforschen, wie anhand von Landschaftsbeschreibungen die Schönheit einer textuell umschriebenen Landschaft geschätzt werden kann. Methoden des Opinion Minings wurden auf die Landschaftsevaluierung angewendet. Die dafür verwendeten Daten enthalten eine grosse Anzahl Beschreibungstexte, die zu Fotografien geschrieben wurden und Bewertungen der Landschaftsqualität auf der Grundlage dieser Bilder. Die eine Hälfte der Daten wurde zur Erstellung verschiedener Lexika verwendet, deren Leistungsfähigkeit an der anderen Hälfte validiert wurde. Mit der Auswertung dieser Daten wurde Folgendes erreicht:

- Nach der Aufarbeitung des wissenschaftlichen Hintergrundes wurde erörtert, welche Auswirkungen dieser auf die verwendeten Daten hat.
- Mit einer explorativen Datenanalyse wurden die Strukturen in den verwendeten Datensätzen aufgedeckt. Die wesentlichen Produkte dieser Analyse sind Grafiken, die die multidimensionalen Skalierungen visualisieren, und die Wortwolken.
- Aufgrund der explorativen Datenanalyse wurde die Untauglichkeit eines generischen Stimmungslexikons für die gestellte Aufgabe dargelegt. Deshalb wurde mit den vorhandenen Daten spezifisch auf diese Aufgabe zugeschnittene Schönheitslexika gebildet.
- Die erstellten Lexika umfassen verschiedene Unigrammlexika, ein Bi- und ein Trigrammlexikon sowie ein Kookkurrenzlexikon. Die Genauigkeit der Ästhetikschätzung mit diesen verschiedenen Lexika wurde evaluiert.
- Die Texte wurden durch ein selbst geschriebenes Programm analysiert und die einzelnen Bestandteile mit den Lexika in der Datenbank abgeglichen. Zusätzlich filtert ein Toponymerkennungsmodul Ortsnamen aus den Texten. Durch die Datenbankabfrage kann für jeden Text ein Schätzwert der Ästhetik berechnet werden, der mit dem durch menschliche Bewerter vergebenen Wert verglichen werden kann.
- Die besten Resultate wurden mit einem Unigrammlexikon bei langen Texten erreicht. Hier konnte mehr als die Hälfte der Varianz der Landschaftsbewertungen durch das implementierte Modell erklärt werden.

– Mit der Schätzung anhand der Bildlegenden von 280'000 bisher unbewerteten Bildern konnte das räumliche Verteilungsmuster der Landschaftsbewertungen in Grossbritannien nachgebildet werden.

Der Beitrag dieser Arbeit ist, dass gezeigt wurde, wie zur flächendeckenden Evaluierung der Ästhetik von Landschaften neue textuelle Datenquellen benutzt werden können, die bisher nicht für die Landschaftsevaluierung verwendet wurden. Ausserdem wurde demonstriert, dass Text Mining einen neuen Blick auf die Wahrnehmung von Landschaften ermöglicht.

8.2 Erkenntnisse

Der Autor weiss von keiner anderen Arbeit, die Methoden des Opinion Mining auf Landschaftsevaluierungen angewendet hat. Direkte Vergleiche mit bestehender Literatur sind daher nicht möglich. Die Resultate haben aber deutlich gezeigt, dass es möglich ist, die beobachteten räumlichen Muster der Landschaftsästhetik mit der Auswertung der Texte nachzumodellieren.

Landschaftsschönheitslexikon

Durch die explorative Analyse konnte die intuitive Vermutung des höheren ästhetischen Wertes von naturräumlichen Landschaften bestätigt werden, weil dieser Zusammenhang auch in den textuellen Daten gefunden wurde. Aufgrund dieser Erkenntnis konnte durch Schwarmintelligenz ein umfangreiches und leistungsstarkes landschaftliches Schönheitslexikon erstellt werden kann, insbesondere wenn man sich auf Unigramme beschränkt. Obwohl die verwendeten Datensätze sehr gross sind, kann man davon ausgehen, dass zur Bildung eines sehr guten Bi- oder Trigrammlexikons noch ausgiebigere Datensammlungen nützlich wären.

Es konnte gezeigt werden, dass die häufig zu findende Verzerrung der menschlichen Sprache ins Positive auch in den Geograph-Kommentaren gefunden werden kann. Weil aber die negativ bewerteten Kommentare überwiegen, kann geschlossen werden, dass in den benutzten Daten mit einer positiven Sprache über ästhetisch wenig gefällige Landschaften geschrieben wird.

Faktoren zur Verbesserung der Schätzung

Generell verbessert sich die Schätzung des Modells, wenn längere Texte als Input dienen. Ab einer gewissen Länge verschlechtert sich die Schätzung aber wieder. Ausserdem sinkt die Anzahl zur Verfügung stehender Bilder, wenn die Anforderungen an die Kommentarlänge

steigen. Weiter hat sich gezeigt, dass es wichtiger ist, möglichst viele Begriffe zur Schätzung zur Verfügung zu haben, anstatt weniger, dafür durch viele Stimmen sehr sicher bewertete Begriffe einzusetzen. Bereits mit einem kleinen Bestand von sehr häufigen Worten konnten aber erstaunlich präzise Schätzungen erstellt werden.

Wenn Toponyme nicht herausgefiltert werden, verbessert sich die Schätzung leicht, aber um die Leistungsfähigkeit des Sprachmodells zu testen, wurde versucht, die Toponyme nicht zu Hilfe zu nehmen. Wenn anstatt der Reproduktion der Bewertungsnote nur die Rangreihenfolge zur Evaluierung der Güte des Modells betrachtet wird, verbessert sich die Korrelation weiter.

Variation der Schönheitswerte

Durch die Variation der Stimmungswerte je nach Ort konnte keine Verbesserung der Schätzung erreicht werden. Es ist deshalb unklar, wie sich die Schönheitswerte regional unterscheiden können und ab welcher räumlichen Auflösung signifikante Unterschiede auftreten. Es ist unbekannt, inwiefern die ermittelten Werte universell gültig sind oder ob sich diese kulturell unterscheiden.

8.3 Ausblick

In weiterführenden Forschungsarbeiten könnte versucht werden, mit Methoden des *natural language processing* die Ästhetik von grossflächigen Gebieten mit anderen Datenquellen vorherzusagen. Als Datenquellen könnten beispielsweise Datensätze mit nicht landschaftlich bewerteten Fotografien (insbesondere Flickr) verwendet werden. Anstatt ganzen Sätzen können solche Datensätze auch nur Stichworte enthalten, was eine zusätzliche Herausforderung darstellt. Eine weitere Datenquelle könnten geografisch verortete Texte sein, die unabhängig von Bildern existieren. Für die Behandlung solcher Fragen könnte das in dieser Arbeit entwickelte Unigrammlexikon verwendet werden.

Wie bereits erwähnt wurde, basiert das entwickelte Stimmungslexikon auf Daten aus Grossbritannien. Weitere Forschungsarbeiten könnten prüfen, inwiefern sich die hier ermittelten Stimmungswerte auf andere Gebiete übertragen lassen und falls es Unterschiede gibt, ab welcher räumlichen Auflösung diese auftreten.

Als mögliche Anwendung könnten Werbetexte für Tourismusdestinationen mit den beschriebenen Methoden analysiert werden, damit sie für ein möglichst positives Bild des Reisezieles optimiert werden können.

9 Anhang

9.1 Liste der häufigen englischen Wörter

'tis	dear	however	nor	their	when'd
'twas	did	I	not	them	when'll
a	didn't	I'd	of	then	when's
able	do	I'll	off	there	where
about	does	I'm	often	there's	where'd
across	doesn't	I've	on	these	where'll
after	don't	if	only	they	where's
ain't	either	in	or	they'd	which
all	else	into	other	they'll	while
almost	ever	is	our	they're	who
also	every	isn't	own	they've	who'd
am	for	it	rather	this	who'll
among	from	it's	said	tis	who's
an	get	its	say	to	whom
and	got	just	says	too	why
any	had	least	shan't	twas	why'd
are	has	let	she	us	why'll
aren't	hasn't	like	she'd	wants	why's
as	have	likely	she'll	was	will
at	he	may	she's	wasn't	with
be	he'd	me	should	we	won't
because	he'll	might	should've	we'd	would
been	he's	might've	shouldn't	we'll	would've
but	her	mightn't	since	we're	wouldn't
by	hers	most	so	were	yet
can	him	must	some	weren't	you
can't	his	must've	than	what	you'd
cannot	how	mustn't	that	what'd	you'll
could	how'd	my	that'll	what's	you're
could've	how'll	neither	that's	when	you've
couldn't	how's	no	the	when	your

Index

Literatur

Anderson, L., und H. Schroeder (1983), Application of wildland scenic assessment methods to the urban landscape, *Landscape Planning*, *10*(3), 219–237.

Arriaza, M., J. Canas-Ortega, J. Canas-Madueno, und P. Ruiz-Aviles (2004), Assessing the visual quality of rural landscapes, *Landscape and urban planning*, *69*(1), 115–125.

Arthur, L., T. Daniel, und R. Boster (1977), Scenic assessment: An overview, *Landscape planning*, *4*, 109–129.

Athar, A. (2011), Sentiment analysis of citations using sentence structure-based features, in *Proceedings of the ACL*, pp. 81–87.

Baccianella, S., A. Esuli, und F. Sebastiani (2010), Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Language Resources Association (ELRA)*.

Benfield, J., P. Bell, L. Troup, und N. Soderstrom (2010), Aesthetic and affective effects of vocal and traffic noise on natural landscape assessment, *Journal of environmental psychology*, *30*(1), 103–111.

Boucher, J., und C. Osgood (1969), The pollyanna hypothesis, *Journal of Verbal Learning and Verbal Behavior*, *8*(1), 1–8.

Bruce, R., und J. Wiebe (1999), Recognizing subjectivity: A case study in manual tagging, *Natural Language Engineering*, *5*(2), 187–205.

Buhyoff, G., und J. Wellman (1979), Seasonality bias in landscape preference research, *Leisure Sciences*, *2*(2), 181–190.

Buhyoff, G., R. Hull, J. Lien, und H. Cordell (1986), Prediction of scenic quality for southern pine stands, *Forest Science*, *32*(3), 769–778.

Carbonell, J. (1979), Subjective understanding: Computer models of belief systems., *Tech. rep.*, DTIC Document.

Cherem, G. (1973), Looking through the eyes of the public, in *Proceedings of Aesthetics Opportunity Colloquium*, pp. 52–64.

Craik, K. H. (1972a), Psychological factors in landscape appraisal, *Environment and Behavior*, *4*(3), 255–266.

Craik, K. H. (1972b), Appraising the objectivity of landscape dimensions, *Natural environments: Studies in theoretical and applied analysis*, *292*.

Dakin, S. (2003), There's more to landscape than meets the eye: Towards inclusive landscape assessment in resource and environmental management, *Canadian Geographer/Le Géographe canadien*, *47*(2), 185–200.

Daniel, T. (2001), Whither scenic beauty? Visual landscape quality assessment in the 21st century, *Landscape and urban planning*, *54*(1), 267–281.

Daniel, T., R. Boster, und R. Forest (1976), *Measuring landscape esthetics: The scenic beauty estimation method*, Rocky Mountain Forest and Range Experiment Station Fort Collins, CO.

Dave, K., S. Lawrence, und D. M. Pennock (2003), Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528, ACM.

Esuli, A., und F. Sebastiani (2006), Sentiwordnet: A publicly available lexical resource for opinion mining, in *Proceedings of LREC*, Vol. 6, pp. 417–422.

Flanagin, A., und M. Metzger (2008), The credibility of volunteered geographic information, *GeoJournal*, *72*(3), 137–148.

Fortuna, B., D. Mladenić, und M. Grobelnik (2005), Visualization of text document corpus.

Fredrickson, L., und D. Anderson (1999), A qualitative exploration of the wilderness experience as a source of spiritual inspiration, *Journal of environmental psychology*, *19*(1), 21–39.

Garcia, D., A. Garas, und F. Schweitzer (2011), Positive words carry less information than negative words, *Arxiv preprint arXiv:1110.4123*.

Gelbukh, A., S. Levachkine, und S.-Y. Han (2004), Resolving ambiguities in toponym recognition in cartographic maps, in *Graphics Recognition. Recent Advances and Perspectives*, pp. 75–86, Springer.

Geograph (2012), http://www.geograph.org.uk, Zugriff am 7.12.2012.

Hatzivassiloglou, V., und K. McKeown (1997), Predicting the semantic orientation of adjectives, in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp. 174–181, Association for Computational Linguistics.

Heintzman, P. (2008), Men's wilderness experience and spirituality: Further explorations, in *Proceedings of the 2007 Northeastern Recreation Research Symposium*, pp. 55–59.

Henwood, K., und N. Pidgeon (2001), Talk about woods and trees: Threat of urbanization, stability, and biodiversity, *Journal of Environmental Psychology*, *21*(2), 125–147.

Herzog, T. (1985), A cognitive analysis of preference for waterscapes, *Journal of Environmental Psychology*, *5*(3), 225–241.

Hill, S., und N. Ready-Campbell (2011), Expert stock picker: The wisdom of (experts in) crowds, *International Journal of Electronic Commerce*, *15*(3), 73–102.

Hodge, V. J., und J. Austin (2002), A comparison of a novel neural spell checker and standard spell checking algorithms, *Pattern Recognition*, *35*(11), 2571–2580.

Hodgson, R., und R. Thayer (1980), Implied human influence reduces landscape beauty, *Landscape Planning*, *7*(2), 171–179.

Hollenstein, L., und R. Purves (2012), Exploring place through user-generated content: Using flickr tags to describe city cores, *Journal of Spatial Information Science*, (1), 21–48.

Hudson, R. (1994), About 37% of word-tokens are nouns, *Language*, pp. 331–339.

Hull, R., und M. McCarthy (1988), Change in the landscape, *Landscape and Urban Planning*, *15*(3), 265–278.

Hull, R., G. Revell, et al. (1989), Issues in sampling landscapes for visual quality assessments, *Landscape and Urban Planning*, *17*(4), 323–330.

Hunziker, M., und F. Kienast (1999), Potential impacts of changing agricultural activities on scenic beauty – a prototypical technique for automated rapid assessment, *Landscape Ecology*, *14*(2), 161–176.

Jackson, J. (1986), The vernacular landscape, *Landscape meanings and values*, pp. 48–64.

Kisilevich, S., C. Rohrdantz, und D. Keim (2010), "Beautiful picture of an ugly place". Exploring photo collections using opinion and sentiment analysis of user comments, in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, pp. 419–428, IEEE.

Kukich, K. (1992), Techniques for automatically correcting words in text, *ACM Computing Surveys (CSUR)*, *24*(4), 377–439.

Lambek, J. (1958), The mathematics of sentence structure, *The American Mathematical Monthly*, *65*(3), 154–170.

Laurie, I. (1974), *Aesthetic Factors in Visual Evaluation*, University of Manchester, Landscape Evaluation Research Project.

Lothian, A. (1999), Landscape and the philosophy of aesthetics: Is landscape quality inherent in the landscape or in the eye of the beholder?, *Landscape and urban planning*, *44*(4), 177–198.

Lund, K., und C. Burgess (1996), Producing high-dimensional semantic spaces from lexical co-occurrence, *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Malm, W., K. Kelley, J. Molenar, und T. Daniel (1981), Human perception of visual air quality (uniform haze), *Atmospheric Environment (1967)*, *15*(10), 1875–1890.

Matsuo, Y., und M. Ishizuka (2004), Keyword extraction from a single document using word co-occurrence statistical information, *International Journal on Artificial Intelligence Tools*, *13*(01), 157–169.

McNaught, C., und P. Lam (2010), Using Wordle as a supplementary research tool, *The qualitative report*, *15*(3), 630–643.

Michel, J.-B., et al. (2011), Quantitative analysis of culture using millions of digitized books, *science*, *331*(6014), 176–182.

Mohammad, S., C. Dunne, und B. Dorr (2009), Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 599–608, Association for Computational Linguistics.

Morinaga, S., K. Yamanishi, K. Tateishi, und T. Fukushima (2002), Mining product reputations on the web, in *Conference on Knowledge Discovery in Data: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Vol. 23, pp. 341–349, Citeseer.

Nagy, W. E., und R. C. Anderson (1984), How many words are there in printed school english?, *Reading Research Quarterly*, pp. 304–330.

Nasukawa, T., und J. Yi (2003), Sentiment analysis: Capturing favorability using natural language processing, in *Proceedings of the 2nd international conference on Knowledge capture*, pp. 70–77, ACM.

Nielsen, J. (2006), Participation inequality: Encouraging more users to contribute, *Jakob Nielsen's alertbox*, *9*, 2006.

Ochoa, X., und E. Duval (2008), Quantitative analysis of user-generated content on the web, in *Proceedings of webevolve2008: Web science workshop at WWW2008*, pp. 1–8.

Pak, A., und P. Paroubek (2010), Twitter as a corpus for sentiment analysis and opinion mining, in *Proceedings of LREC*, Vol. 2010.

Pang, B., und L. Lee (2008), *Opinion Mining and Sentiment Analysis*, Now Pub.

Perrie, J., A. Islam, E. Milios, und V. Keselj (2013), Using google n-grams to expand word-emotion association lexicon, in *Computational Linguistics and Intelligent Text Processing*, pp. 137–148, Springer.

Piantadosi, S., H. Tily, und E. Gibson (2011), Word lengths are optimized for efficient communication, *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Polanyi, L., und A. Zaenen (2006), Contextual valence shifters, *Computing attitude and affect in text: Theory and applications*, pp. 1–10.

Popescu, A., und O. Etzioni (2005), Extracting product features and opinions from reviews, in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 339–346, Association for Computational Linguistics.

Purves, R., A. Edwardes, und J. Wood (2011), Describing place through user generated content, *First Monday*, *16*(9-5).

Roth, M. (2006), Validating the use of internet survey techniques in visual landscape assessment—an empirical study from Germany, *Landscape and Urban Planning*, *78*(3), 179–192.

Rozin, P., L. Berman, und E. Royzman (2010), Biases in use of positive and negative words across twenty natural languages, *Cognition and Emotion*, *24*(3), 536–548.

Sack, W. (1995), Representing and recognizing point of view, in *Proc. AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*.

Samet, M. D. L. H. (2011), Multifaceted toponym recognition for streaming news.

ScenicOrNot (2012), http://scenic.mysociety.org, Zugriff am 10.12.2012.

Schroeder, H. (1991), Preference and meaning of arboretum landscapes: Combining quantitative and qualitative data, *Journal of Environmental Psychology*, *11*(3), 231–248.

Schroeder, H., und T. Daniel (1981), Progress in predicting the perceived scenic beauty of forest landscapes, *Forest Science*, *27*(1), 71–80.

Shafer, E., und J. Mietz (1970), *It Seems Possible to Quantify Scenic Beauty in Photographs*, Vol. 162, US Northeastern Forest Experiment Station.

Shuttleworth, S. (1979), The evaluation of landscape quality, *Landscape Research*, *5*(1), 14–15.

Siersdorfer, S., E. Minack, F. Deng, und J. Hare (2010), Analyzing and predicting sentiment of images on the social web, in *Proceedings of the international conference on Multimedia*, pp. 715–718, ACM.

Stadler, B. (2010), Zusammenhänge zwischen Bildbewertungen und Landschaftsklassen - Erklärungsgehalt der CORINE Land Cover-Klassen für die subjektiven Schönheitsbewertungen in einer Bilddatenbank in Grossbritannien, Master's thesis, Geographisches Institut Universität Zürich.

Stamps, A. E. (1990), Use of photographs to simulate environments: A meta-analysis, *Perceptual and Motor Skills*, *71*(3), 907–913.

Stewart, T., P. Middleton, M. Downton, und D. Ely (1984), Judgments of photographs vs. field observations in studies of perception and judgment of the visual environment, *Journal of Environmental Psychology*, *4*(4), 283–302.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, und M. Stede (2011), Lexicon-based methods for sentiment analysis, *Computational Linguistics*, *37*(2), 267–307.

Terveen, L., W. Hill, B. Amento, D. McDonald, und J. Creter (1997), Building task-specific interfaces to high volume conversational data, in *Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI conference on Human factors in computing systems*, Vol. 22, pp. 226–233.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, und A. Kappas (2010), Sentiment strength detection in short informal text, *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558.

Thelwall, M., K. Buckley, und G. Paltoglou (2012), Sentiment strength detection for the social web, *Journal of the American Society for Information Science and Technology*.

Tobler, W. R. (1970), A computer movie simulating urban growth in the Detroit region, *Economic geography*, *46*, 234–240.

Tuan, Y. (1979), Thought and landscape, *The interpretation of ordinary landscapes. Oxford University Press, New-York*.

Tulloch, D. (2007), Many, many maps: Empowerment and online participatory mapping, *First Monday*, *12*(2), 5.

Turney, P. D. (2002), Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424, Association for Computational Linguistics.

Van den Berg, A. E., C. A. Vlek, und J. F. Coeterier (1998), Group differences in the aesthetic evaluation of nature development plans: A multilevel approach, *Journal of environmental psychology*, *18*(2), 141–157.

Viegas, F. B., M. Wattenberg, und J. Feinberg (2009), Participatory visualization with Wordle, *Visualization and Computer Graphics, IEEE Transactions on*, *15*(6), 1137–1144.

Wherrett, J. (1999), Issues in using the internet as a medium for landscape preference research, *Landscape and Urban Planning*, *45*(4), 209–217.

Whitelaw, C., N. Garg, und S. Argamon (2005), Using appraisal groups for sentiment analysis, in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 625–631, ACM.

Wiebe, J. (2000), Learning subjective adjectives from corpora, in *Proceedings of the National Conference on Artificial Intelligence*, pp. 735–741, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Wiebe, J., T. Wilson, und C. Cardie (2005), Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation*, *39*(2), 165–210.

Wilks, Y., und J. Bien (1983), Beliefs, points of view, and multiple environments, *Cognitive Science*, *7*(2), 95–119.

Wise, J. A. (1999), The ecological approach to text visualization, *Journal of the American Society for Information Science*, *50*(13), 1224–1233.

Yi, J., T. Nasukawa, R. Bunescu, und W. Niblack (2003), Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques, in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 427–434, IEEE.

Zipf, G. (1935), The psycho-biology of language, *Houghton Mifflin, Oxford*.

Persönliche Erklärung:
Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Mario Nowak