WaMos 3: Aktualisierung Social Media-Analyse

Ross Purves (ross.purves@geo.uzh.ch), Manuel Baer, Maximilian Hartmann und

Sharon Richardson

Geographisches Institut, Universität Zürich

8. Dezember 2020

1. Einführung

Der vorliegende Bericht beschreibt die Ergebnisse eines Projekts zur Aktualisierung einiger Messungen der Aktivitäten im Schweizer Wald mittels Social Media. Die ursprüngliche Projektbeschreibung sah eine Wiederholung unserer ursprünglichen Arbeit, über die in Wartmann et al. (2018) berichtet wurde, durch die Charakterisierung der Erholungswaldnutzung in der Schweiz mit Hilfe von Social Media-Inhalten vor. In der Pilotstudie untersuchten und verglichen wir drei Quellen:

- Flickr
- Instragram
- Twitter

Aufgrund von Änderungen in der Art und Weise, wie auf diese Datenquellen zugegriffen werden kann, und im Hinblick auf die Geolokalisierung von Inhalten, konzentriert sich dieser Abschlussbericht, wie nach unserem Zwischenbericht vereinbart, in seiner Analyse auf die Verwendung einer sozialen Medienform, nämlich Flickr-Beiträge.

In diesem Forschungsbericht haben wir:

1. die bisherige Analyse der potenziellen Erholungsnutzung des Waldes in der Schweiz mit einem aktuellen Datensatz wiederholt;
2. die Muster der Erholungsnutzung der Schweiz über biogeografische Regionen und Jahreszeiten untersucht; und
3. die räumliche und zeitliche Variation in der Nutzung von Erholungswald durch eine Textanalyse untersucht, wobei der Schwerpunkt auf Landschaftselementen, Qualitäten und Aktivitäten lag

Der Bericht ist wie folgt strukturiert. Zunächst stellen wir kurz die wichtigsten Ideen im Hinblick auf die Analyse von Social Media vor und präsentieren einen allgemeinen Rahmen für die Analyse von Social Media-Daten (vgl. Wartmann et al. 2020). Dann stellen wir die aktuelle Situation in Bezug auf den Zugang zu Social Media-Daten dar und erläutern unseren Entscheid in dieser Studie nur mit Flickr-Daten als Datenquelle zu arbeiten. Wir stellen kurz den vollständigen Flickr-Datensatz vor, mit dem wir gearbeitet haben, und erläutern die wichtigsten Filterschritte, die als Ausgangspunkt für die oben aufgeführten Fallstudien durchgeführt wurden. Wir präsentieren und interpretieren jede Fallstudie einzeln. Der Bericht schliesst mit einer Erörterung des Potenzials der sozialen Medien als Informationsquelle im Hinblick auf die Erholungsnutzung des Waldes und einigen Empfehlungen für mögliche zukünftige Arbeiten.

2. Analysieren von Social Media-Daten

Die Nutzung von Social Media als Datenquelle in der geographischen Forschung hat sich in den letzten 15 Jahren zu einem sehr aktiven Forschungsgebiet entwickelt. Die Nutzung von Social Media umfasst vielfältige Themen, die von politischen Debatten und ihrer Entwicklung über Gesundheitsgeographie und die Charakterisierung von Naturkatastrophen bis hin zur Verwendung von Ortsnamen und der

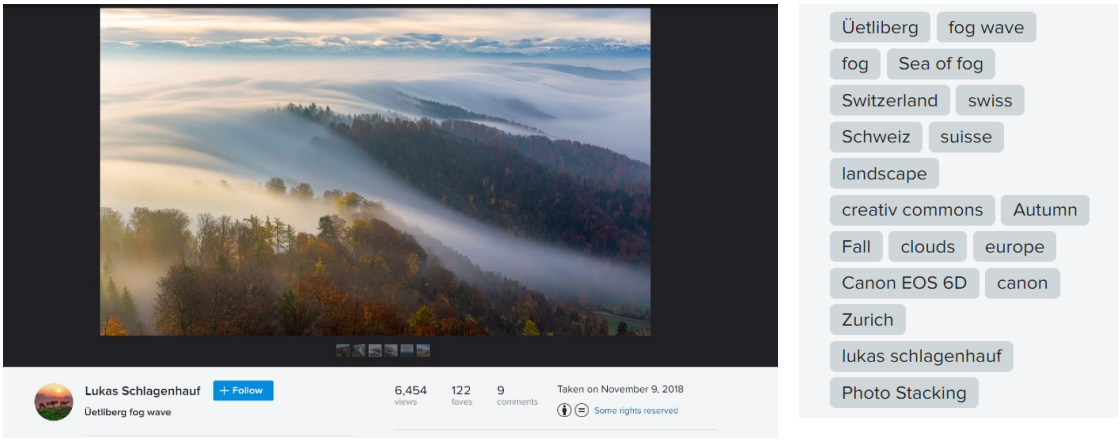
Erforschung von Landschaftspräferenzen[1] reichen. Im Mittelpunkt all dieser Studien stehen Fragen, bei denen der Standort für die untersuchte Frage wichtig ist. So haben Forscher*innen zum Beispiel untersucht, ob positive oder negative Empfindungen in Bezug auf das Brexit Referendum mit den Ergebnissen von Volksabstimmungen an bestimmten Orten zusammenhängen. Andere Forschungsgruppen haben zum Beispiel Regionen aufgrund ihrer umgangssprachlichen Namen, welche aus online Dokumenten extrahiert wurden, identifizieren können. Solche Studien stützen sich auf geolokalisierte soziale Medien - d.h. die Verknüpfung eines Social Media-Posts mit dem physischen Standort. Diese Verknüpfung kann entweder explizit - in Form von Koordinaten oder der eindeutigen Verwendung von Ortsnamen, oder implizit - beispielsweise durch die Erkennung der Verwendung von Wörtern, die mit bestimmten Orten assoziiert sind, erfolgen. In unserer Arbeit analysieren wir explizit georeferenzierte Inhalte - d.h. Social Media-Posts, bei denen dem Inhalt Koordinaten hinzugefügt wurden. Diese Koordinaten werden in der Regel auf eine von zwei Arten hinzugefügt:

- Automatisch durch das Gerät, das den Inhalt postet (z.B. werden Koordinaten mit Hilfe des GPS eines mobilen Geräts hinzugefügt, wenn ein Foto gemacht wird)
- Indem ein Benutzer beim Hochladen von Inhalten geografische Informationen hinzufügt, z.B. einen Ortsnamen hinzufügt oder ein Fotoalbum auf einer Karte lokalisiert

Neben dem Standort ermöglichen es uns Social Media-Posts oft auch, Inhalte mit bestimmten Zeiten und eindeutigen Benutzern zu verknüpfen. Wir sprechen hier von einem Datenatom *a:*

$$a = \{u, l, t, c\}$$

wobei *u* der Benutzer, *l* der Standort, *t* eine Zeit und *c* der gepostete Inhalt ist.



a= {Lukas Schlagenhauf, (8.491388; 47.349444), 9.11.2018, (Bild; Tags; Kommentare; Likes)}

Abbildung 1: Beispiel eines Social Media-Posts, der potenzielle Dimensionen der Analyse zeigt (Inhalte unter einer CC-BY-Lizenz unter https://www.flickr.com/photos/lschlagenhauf/45126137264/ veröffentlicht)

Abbildung 1 illustriert ein Beispiel für ein solches einzelnes Datenatom anhand eines Flickr-Beitrags, das von einem namentlich genannten Benutzer im Herbst 2018 aufgenommen und mit geografischen Koordinaten verknüpft wurde. Die Koordinaten platzieren das Bild auf dem Gipfel des beliebten Erholungswaldgebietes auf dem Uetliberg in der Nähe von Zürich. Unser Ansatz zur Analyse von Social Media-Daten kombiniert verschiedene Aspekte solcher Datenatome, um aussagekräftige

[1] Für eine Literaturübersicht zur Nutzung von Social Media in der Freizeitforschung verweisen wir auf Wartmann et al. (2018).

Gesamtinformationen zu extrahieren. Dabei folgen wir einem standardisierten generischen Rahmen für die Analyse von Social Media-Daten (vgl. Purves & Mackaness 2016; Toivonen et al. 2019; Wartmann et al. 2020). Dieser Rahmen wird im Folgenden vorgestellt und wurde bei der Entwicklung der folgenden drei Fallstudien jeweils verwendet.

Rahmen für die Analyse sozialer Medien

1. Forschungsfrage identifizieren

Bevor mit der Analyse begonnen werden kann, ist es unerlässlich, die zu analysierende Forschungsfrage zu diskutieren. Die Spezifizierung der Frage hilft, geeignete potenzielle Social Media-Datenquellen und -methoden zu identifizieren, und ist auch wichtig, um potenzielle Verzerrungen in den Daten und deren Handhabung zu berücksichtigen.

2. Sammlung von Daten

Die Datenerhebung kann vier grundlegende Formen annehmen:

1) In einer kleinen Anzahl von Fällen sind vollständige Datensätze unter Creative-Commons-Lizenzen verfügbar (z.B. www.geograph.org.uk).
2) Einige Social Media-Plattformen (z. B. Flickr und Twitter) bieten den Zugriff auf ihre Daten über Application Programming Interfaces (APIs) an. Dieser Zugang kann kostenlos sein, und die API erlaubt in der Regel die Angabe einer Reihe von Suchparametern. Beispielsweise kann Flickrs gesamtes Archiv anhand von Schlüsselwörtern oder geografischen Begrenzungskästen durchsucht werden. Im Gegensatz dazu kann Twitter anhand von Parametern durchsucht werden, jedoch nur nach aktuellen Daten.
3) In einigen Fällen sind historische Daten auch über kostenpflichtige Dienste und Abonnements verfügbar. Somit ist der Zugang zu einer breiteren Palette von Daten möglich, im Vergleich zu den jeweiligen öffentlichen APIs.
4) Schliesslich ist es möglich, Daten mit Hilfe von selber geschriebenem Code zu «scrapen». Dabei werden von einer online Ressource, mithilfe eines Programms, die erwünschten Daten extrahiert. Dieser Ansatz wird zwar häufig verwendet, verstösst jedoch häufig explizit gegen die von den Anbietern festgelegten Bedingungen und Konditionen.

3. Datenfilterung für Verzerrung

Eine Schlüsselfrage in Bezug auf Social Media ist ihre Repräsentativität. Je nach Fragestellung kann es sein, dass diejenigen, die soziale Medien nutzen, nur einen kleinen Teil der untersuchten Bevölkerung ausmachen und wichtige Gruppen vernachlässigt werden (z.B. Kinder oder Rentner*innen). Ebenso wichtig zu berücksichtigen sind Verzerrungen, die den sozialen Medien inhärent sind, und solche, die nur auf spezifischen Plattformen zu finden sind. Zu den inhärenten Verzerrungen gehört die Ungleichheit der Beteiligung - dass ein kleiner Teil der Gesamtzahl der Nutzer die grösste Datenmengen erzeugt (Haklay 2016). Ein Beispiel für plattformspezifische Verzerrungen sind Massen-Uploads in Flickr - ein Phänomen, bei dem ein Benutzer eine grosse Anzahl von Bildern mit identischen Tags und/oder Koordinaten beisteuert, was manchmal zu Hunderten von Bildern mit denselben Metadaten führt. Dieses Verhalten kann so interpretiert werden, dass anstelle von einzelnen Bildern mit individuellen Metadaten, ein Album mit Bildern (z.B. von einer Urlaubsreise) angelegt wird. Die Berücksichtigung dieser Art von Verzerrungen ist bei der Analyse von Social Media-Daten wichtig, da sie die Ergebnisse stark beeinflussen können.

4. Relevante Teilmenge von Daten auswählen

Auf der Grundlage der untersuchten Forschungsfrage können wir die Daten unter Verwendung einer oder mehrerer Dimensionen unserer Datenatome weiter filtern. So kann es beispielsweise sinnvoll sein, nur Social Media-Daten innerhalb einer bestimmten Region oder eines bestimmten Landbedeckungstyps zu extrahieren, indem wir den Standort, bestimmte Zeiten oder bestimmte Jahreszeiten oder Daten mit bestimmten Inhalten (z.B. blauer Himmel) oder Tags (z. B. `Wald`) verwenden.

5. Matching

Matching bezieht sich auf den Prozess der Verknüpfung von Daten mit externen Datensätzen. Dies kann trivial sein - zum Beispiel die Verwendung der Koordinaten oder des Zeitstempels, die mit einem Datenatom verbunden sind. In der Praxis ist ein Post jedoch in der Regel nicht mit einem genauen räumlich-zeitlichen Standort verbunden, da er sowohl eine Region als auch ein zeitliches Intervall erfasst (z.B. ist das Bild in Abbildung 1 auf dem Gipfel des Uetlibergs geolokalisiert, zeigt aber tatsächlich eine viel grössere Region und erfasst ein Phänomen (*Hochnebel*), das allgemein mit dem Mittelland in Verbindung gebracht werden kann. Zudem steht dieses Bild zeitlich nicht nur für ein Ereignis an einem bestimmten Tag, sondern vielmehr für ein sich jahreszeitlich wiederholendes meteorologisches Phänomen). Die Wahl eines geeigneten räumlichen und zeitlichen Umfangs für die Zuordnung eines einzelnen Atoms ist daher ein nicht-triviales Problem, das sowohl bei der Verknüpfung von Inhalten verschiedener Formen (z.B. textliche Beschreibungen, die Wanderungen assoziiert mit Flickr-Bildern beschreiben (Wartmann et al. 2018)) als auch bei der Verknüpfung von Inhalten mit Zusatzdaten, die bei der Analyse verwendet werden (z.B. die Verknüpfung von Social Media-Posts mit Points of Interest in OpenStreetMap), von Bedeutung ist.

6. Analyse-Ansätze

Ansätze zur Analyse von Social Media-Daten werden von einer Vielzahl von Faktoren bestimmt. Darunter befinden sich Fragen über die Charakteristiken der einzelnen Datenatome in Bezug auf eine bestimmte Plattform, die verfügbaren Datenmengen, die Zusatzdaten, mit denen die sozialen Medien verknüpft werden können, und vor allem, die untersuchte Forschungsfrage. In der Praxis bedeutet dies, dass das Spektrum der Ansätze zur Analyse von Social Media-Daten sowohl qualitative als auch quantitative Methoden umfasst und typischerweise durch die Hintergründe und die methodische Expertise des Forschungsteams bestimmt wird. Der Schwerpunkt unseres Teams liegt auf skalenbezogenen Raum-Zeitanalysen (z.B. die Identifizierung von Stadtkernen, oder das analysieren von Hotspots beziehungsweise Coldspots, die mit bestimmten Landschaftseigenschaften verbunden sind) und auf der Erforschung einzelner Zielvariablen und ihrer Semantik (z.B. kulturelle Ökosystemdienstleistungen wie Ruhe) (Hollenstein & Purves 2010; Purves & Derungs 2014; Wartmann et al. 2018; Chesnokova et al. 2019).

7. Interpretation und Auswertung

Ein wichtiger letzter Schritt bei jeder Analyse ist die Interpretation der Ergebnisse in Bezug auf die ursprüngliche Forschungsfrage, die Bewertung ihrer Qualität (durch Vergleiche mit anderen Datenquellen wie z.B. traditionellen Fragebogendaten) und Empfehlungen für mögliche weitere Arbeiten auf der Grundlage der Qualität der erzielten Ergebnisse.

3. Änderungen beim Zugang zu Social Media-Daten

In der Pilotstudie verglichen wir die drei Quellen Instagram, Twitter und Flickr. Dieser Vergleich ergab, dass einerseits die aus den drei Quellen verfügbaren Datenmengen sehr unterschiedlich waren, andererseits die Nutzungsmuster zwischen den verschiedenen Quellen aber weitgehend korrelierten. In der Pilotstudie warnten wir davor, dass der Zugang zu kommerziellen Diensten unsicher sei und dass Veränderungen in der Popularität der Plattformen, des Zugangs zu APIs und der Dateneigenschaften selbst die Wiederholung von Studien erschweren könnten. Inzwischen sind an allen drei Quellen massgebende Veränderungen festzustellen, was in zwei Fällen bedeutet, dass sie nicht mehr als Quellen für unsere Arbeit geeignet sind.

1) Nach Bedenken hinsichtlich des Datenschutzes schränkte Instagram den Zugang zu seinem API[2] ein, wodurch der Zugriff auf Bilder und Metadaten für Forschungszwecke und andere Zwecke eingestellt wurde.
2) Auch Twitter reagierte auf Datenschutzbedenken und entfernte im Frühjahr 2019[3] die genaue Georeferenzierung mit Ausnahme der spezifischen Klasse von Bildern (Hu & Wang 2020). Auch wenn Tweets noch grobe Standortsangaben (in der Granularität von Städten und Gemeinden) enthalten können, sind diese im Allgemeinen für unsere Aufgabe nicht mehr geeignet.
3) Flickr wurde verkauft und änderte die Nutzungsbedingungen für einzelne Benutzer. Insbesondere wurde das kostenlose Hosting auf 1000 Bilder (von früher einem Terabyte) begrenzt, es sei denn, die Bilder wurden mit einer Creative-Commons-Lizenz hochgeladen[4].

[2] https://www.instagram.com/developer/

[3] https://twitter.com/TwitterSupport/status/1141039841993355264?s=20

[4] https://techcrunch.com/2018/11/01/flickr-revamps-under-smugmug-with-new-limits-on-free-accounts-unlimited-storage-for-pros/

Diese letzte Bedingung könnte von Interesse sein, da sie die Lizenzierung und Wiederverwendung von Bildern für die Forschung klarer festlegt.

Diese Änderungen bedeuteten jedoch, dass eine Wiederholung unserer früheren Arbeit an Instagram und Twitter nicht mehr möglich war, weswegen wir unsere Analyse in diesem Bericht auf Flickr-Daten beschränken.

4. Übersicht der Fallstudien

Wir führten drei Fallstudien mit Flickr-Daten in einem Schweizer Kontext durch. Alle drei Fallstudien untersuchten Aspekte der Erholungsnutzung des Waldes auf verschiedenen Skalen, die von nationaler bis lokaler Ebene reichten. Unser Ausgangspunkt für diese Studien war ein Datensatz von 3,2 Millionen Flickr-Bilder mit dazugehörigen Metadaten, die zwischen 2007 und 2020 gesammelt wurden. Abbildung 2 zeigt die Anzahl der Bilder pro Monat, die in diesem Zeitraum aus dem API extrahiert wurden. Bemerkenswert ist, dass der Höhepunkt der Flickr-Nutzung um 2014-2015 lag und dass die Nutzung seither auf das Niveau von 2008-2009 zurückgegangen ist. Diese Variation der Popularität von Social Media im Laufe der Zeit, in diesem Fall die Popularität von Flickr, bedeutet, dass die Daten zuerst hinsichtlich der Popularität und somit der Nutzung der Plattform normalisiert werden müssen.

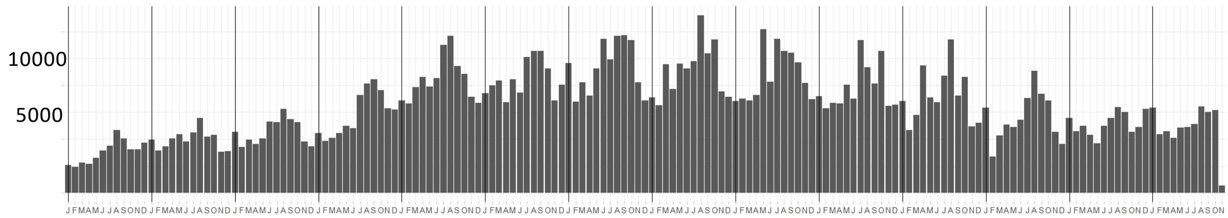


Abbildung 2: Gesamtzahl der von Flickr abgerufenen Flickr-Bilder pro Monat zwischen 2007 und 2020

Die drei von uns durchgeführten Fallstudien werden einzeln beschrieben. Für jede Fallstudie wurden die Metadaten von 3,2 Millionen Bildern, die mithilfe des API's gesammelt wurden, als Ausgangsdaten verwendet. Nachfolgend wird die anschliessende Verarbeitung dieser Rohdaten beschrieben. Unsere Analyse verwendete die folgenden Metadatenelemente:

u: eindeutige anonymisierte Benutzer*innen IDs, die es ermöglichen, Bilder, die von derselben Flickr-Benutzer*innen aufgenommen wurden, zu gruppieren

l: Breiten- und Längengrad-Koordinaten in WGS84, die dem Bild zugeordnet sind

t: ein eindeutiger Zeitstempel, der mit dem Bild verbunden ist. Flickr-Bilder können zwei Zeitstempel haben: einen, der von Geräte-Metadaten abgeleitet ist (die Zeit, zu der das Bild von dem Gerät aufgenommen wurde), und einen, der die Zeit erfasst, zu der das Bild auf Flickr hochgeladen wurde. Da wir nur an groben zeitlichen Informationen interessiert waren, haben wir letztere für unsere Analyse verwendet, da die gerätebezogenen Zeitdaten oft fehleranfällig sind, wenn z.B. die Kameras nicht korrekt eingestellt sind.

c: die einem Bild zugeschriebenen Tags; wir analysierten weder Titel noch Beschreibungen, die mit Bildern verbunden sind.

4.1 Fallstudie eins: Modellierung des Erholungspotentials

Unsere erste Fallstudie wiederholte die Arbeiten der Pilotstudie (Wartmann et al. 2018), in der wir zeigten, dass die Flickr-Daten gut mit dem von der WSL modellierte Freizeitpotenzial (Brändli & Ulmer 2001) korrelierten. Unsere zugrunde liegende Annahme ist, dass Flickr-Bilder häufig Freizeitaktivitäten erfassen und dass ihre Verteilung daher mit der tatsächlichen Erholung und nicht mit dem Erholungspotenzial korreliert. Modellresiduen deuten daher auf potentielle Orte hin, an denen das WSL-Modell vom tatsächlichen (Sozial Medien) Verhalten abweichen könnte.

In einer Modellierungsübung werden Modelle verglichen, die auf dem jetzt verfügbaren vollständigen Datensatz sowie auf einer Auswahl verschiedener zeitlicher Intervalle basieren. In dieser Fallstudie stellen wir die folgende Frage:

> Inwieweit ist das Erholungspotenzial durch die räumliche Variation der Flickr-Daten erfasst? Beziehungsweise, ist die zeitliche Variation des Erholungspotenzials durch die Flickr-Daten erfasst und wie können die Residuen zwischen dem Modell von Brändli & Ulmer (2001) und einem Flickr-basierten Modell verwendet werden, um Regionen für weitere Untersuchungen zu identifizieren?

Wir analysierten das Erholungspotenzial anhand der Positionen der einzelnen Bilder. Daher berücksichtigten wir pro Benutzer nur ein Bild pro eindeutigem Koordinatensatz, wodurch der ursprüngliche Metadatensatz von 3,2 Millionen Bildern auf etwa 2 Millionen reduziert wurde. Alle in der Schweiz gefundenen Flickr-Daten wurden nach der genannten Filterung um Verzerrungen zu reduzieren im Modell berücksichtigt.

Das ursprüngliche Modell von Brändli & Ulmer wurde auf einem Gitter mit einem Punktabstand von 2 km berechnet. Diesen Punkten wurden auch Werte zugeordnet, die angeben, ob sie innerhalb der vom LFI ausgewiesenen Waldfläche liegen oder nicht.

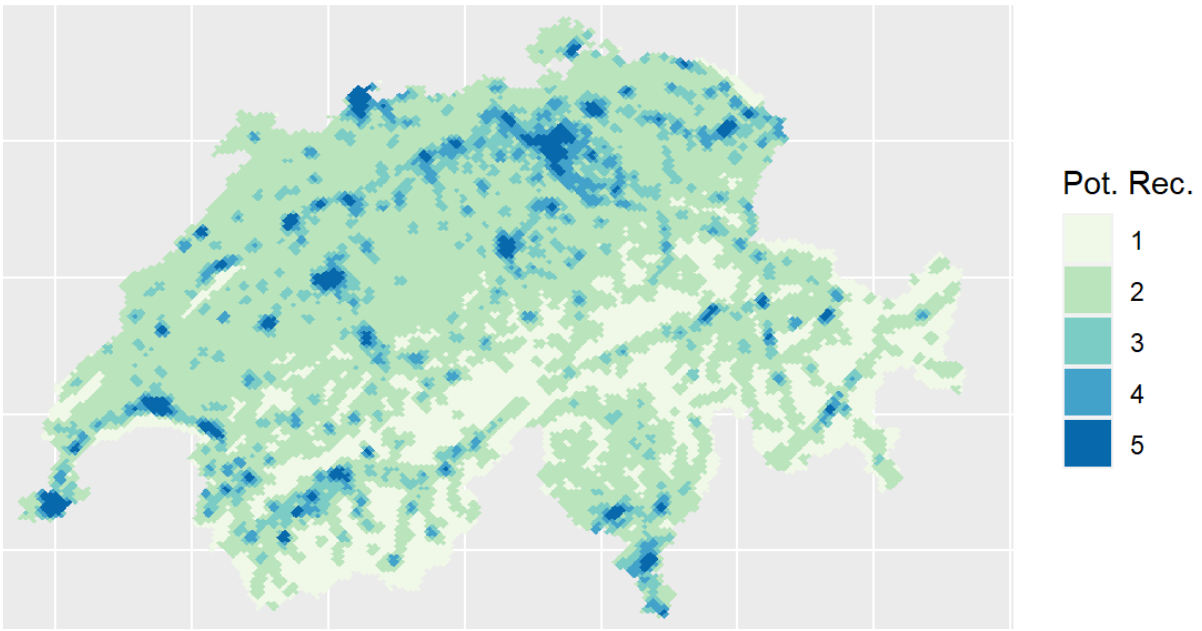


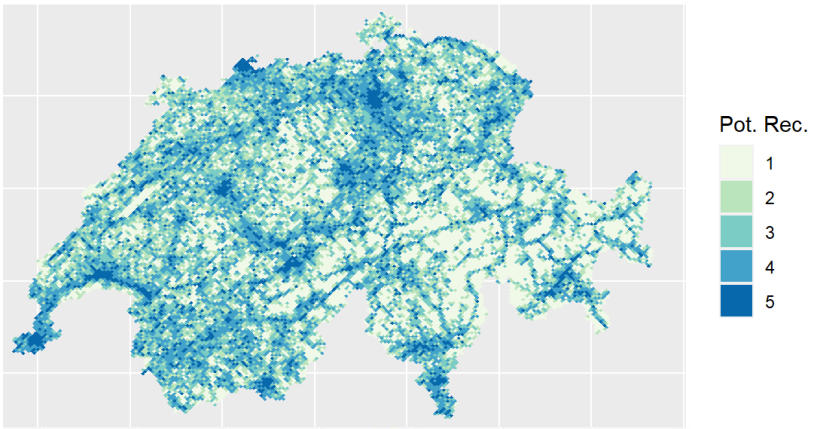
Abbildung 3: Erholungspotenzial nach Brändli & Ulmer (2001)

Unsere Grundannahme war, dass das von Brändli & Ulmer (2001) entwickelte Modell des Erholungspotenzials (Abbildung 3), das eine Reihe von Variablen in Bezug auf Bevölkerungsdichte, Zugänglichkeit und Erholungspotenzial enthält, mit den Standorten der Social Media-Daten korreliert. Dies, weil angenommen wird, dass die Social Media-Daten beliebte Standorte sowohl in städtischen, wie auch in nicht-urbanen Gebieten erfassen.

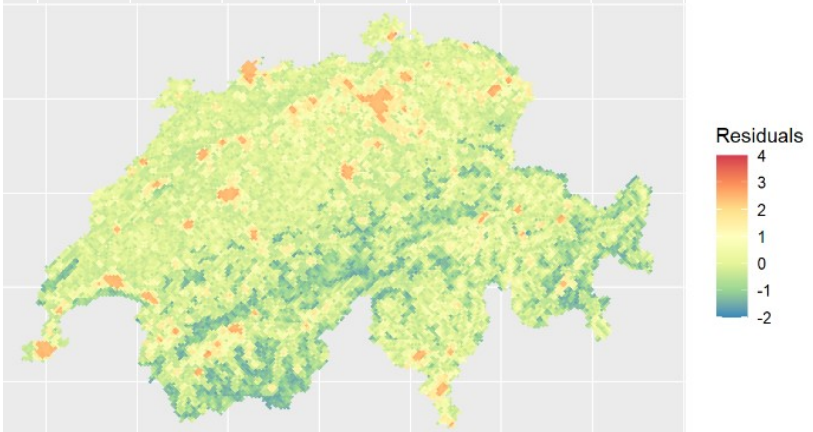
Als Index des Erholungspotenzials berechneten wir den durchschnittlichen Abstand von jedem Erholungspotentialstandortes zu den zehn nächstgelegenen Flickr-Punkten. In einem weiteren Schritt haben wir den Datensatz gefiltert, um nur diejenigen Punkte beizubehalten, welche sich auch in einem LFI-Gebiet befinden.

Um die Beziehung zwischen der Entfernung zu Social Media-Punkten und dem Modell der potenziellen Freizeitgestaltung zu modellieren, verwendeten wir ein *Gerneralised Linear Model* (GLM) (Pinheiro et al. 2020). Da die zugrundeliegenden Daten stark räumlich autokorreliert sind, führten wir auch Modelle durch, die die räumliche Autokorrelation berücksichtigen. Darüber hinaus führten wir individuelle Jahresmodelle durch, um die Variation der Korrelation über die Zeit zu untersuchen.

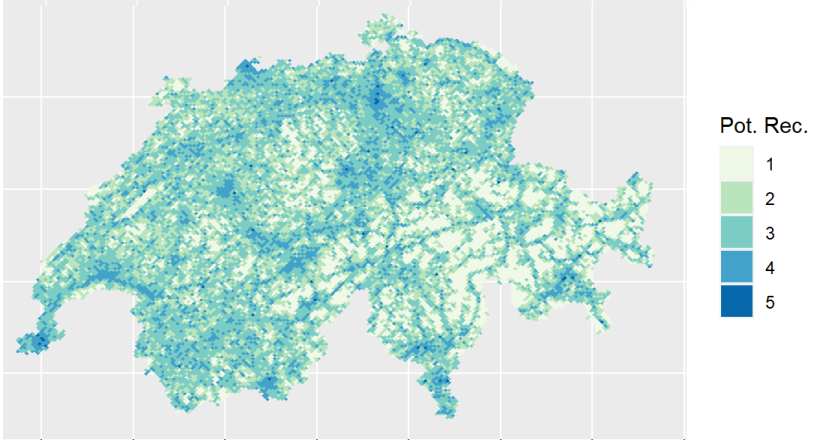
Modellierte potentielle Erholung (GLM)



Residuen (Modell verglichen mit Brändli & Ulmer (2001))



Modellierte potentielle Erholung unter Berücksichtigung der räumlichen Autokorrelation



Residuen unter Berücksichtigung der räumlichen Autokorrelation (Modell verglichen mit Brändli & Ulmer (2001))

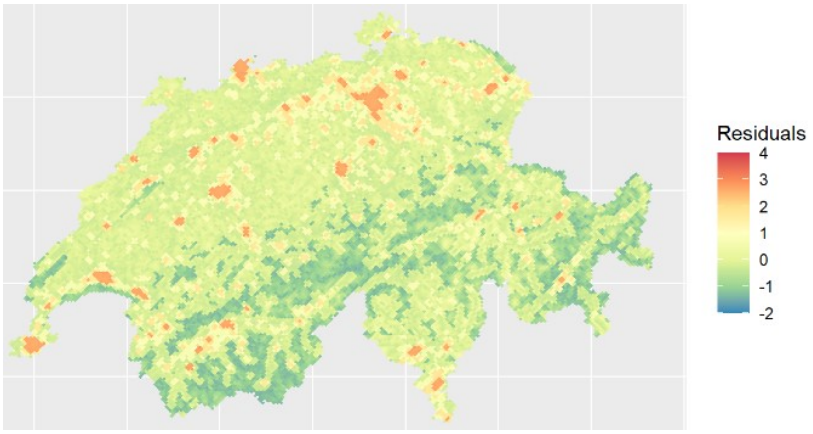
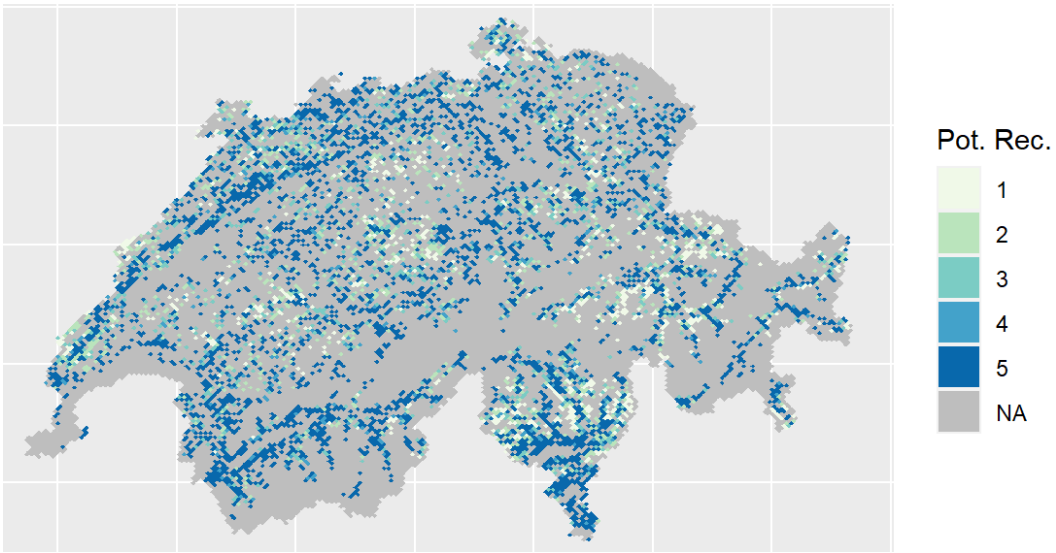


Abbildung 4: Modellierte potentielle Erholung und Residuen

Abbildung 4 zeigt die Ergebnisse der Modellierung des Erholungspotenzials unter Verwendung von Social Media-Daten und die damit verbundenen Residuen. Positive Residuen deuten darauf hin, dass Social Media das Erholungspotenzial nach dem Modell von Brändli & Ulmer unterschätzt, negative Residuen deuten auf eine Überschätzung hin.

Einige Punkte sind hier erwähnenswert. Erstens liefern beide Modelle im Grossen und Ganzen ähnliche Muster, sowohl hinsichtlich des Erholungspotenzials als auch der Residuen. Das Modell, das die räumliche Autokorrelation berücksichtigt, glättet im Wesentlichen das Rauschen im resultierenden Muster. Zweitens unterschätzt Social Media das Erholungspotenzial in städtischen Gebieten, wo Brändli & Ulmer (in ihrem Ordinalmodell) das maximale Potenzial abschätzen, während das von Social Media abgeleitete Modell auf einer kontinuierlichen Skala (niedrigere) Werte liefert. Interessanter ist, dass die Social Media-Daten auf eine Unterschätzung des Erholungspotenzials im Modell von Brändli & Ulmer hinzuweisen scheinen, zum Beispiel im Vallée de Joux in der Westschweiz.

Abbildung 5 zeigt die Ergebnisse, bei denen nur die von der LFI ausgewiesenen Waldflächen beibehalten werden. Hier sind die städtischen Gebiete, die unterschätzt wurden, grösstenteils verschwunden, aber einige wichtige Artefakte (z.B. Vallée de Joux) bleiben erhalten und sind einer weiteren Untersuchung wert.



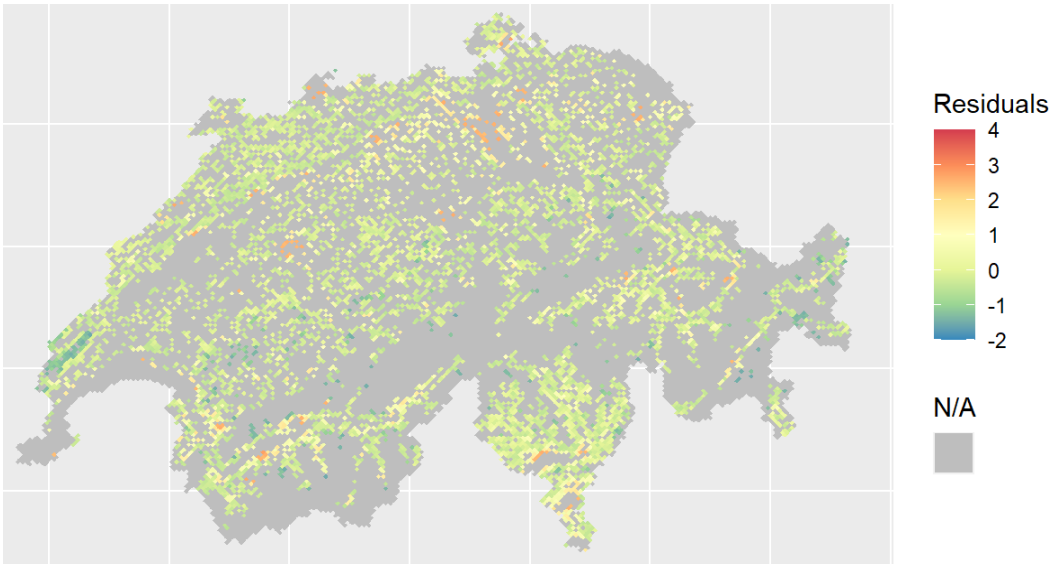


Abbildung 5: Potenzielle Erholungswerte und Residuen für LFI-designierte Waldflächen

Schliesslich fasst Abbildung 6 Residuen zusammen, die auf einer jährlichen Untermenge von Daten, für die von der LFI ausgewiesenen Waldgebiete zwischen 2011 und 2019, basieren. Obwohl kleine Unterschiede im Muster der Residuen sichtbar sind, ist die Beziehung insgesamt im Laufe der Zeit relativ robust. Dies deutet darauf hin, dass jährlich eine ausreichende Anzahl an Fotos auf Flickr hochgeladen wurden, um einen Vergleich mit dem ursprünglichen Modell zu ermöglichen.

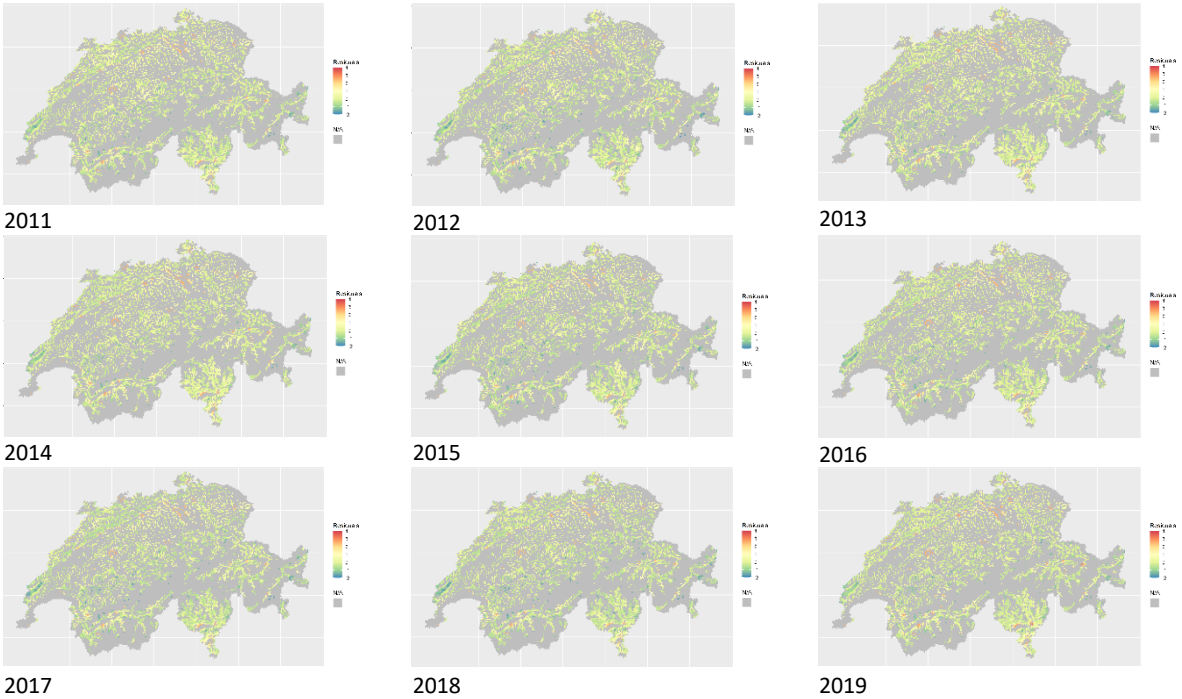


Abbildung 6: Restmengen für das modellierte Erholungspotenzial zwischen 2011 und 2019

4.2 Fallstudie zwei: Räumlich-zeitliche Variation in der Erholungswaldnutzung

Unsere zweite Fallstudie untersucht das Verhalten auf der Ebene der Schweizer Wälder. Die Fragen, die wir in dieser Fallstudie untersuchten, waren

- Auf welchen räumlichen und zeitlichen Skalen können wir Muster beim Besuch der Schweizer Wälder mit Hilfe von Social Media-Daten untersuchen, und was verraten uns diese Muster über das Verhalten der Besucher?

Da wir uns für die Anzahle der Besuche der Waldgebiete interessierten, extrahierten wir zunächst alle Flickr-Bilder, die innerhalb von Waldparzellen gefunden wurden. Die Waldparzellen sind in SwissTopos TLM Regio digitalisiert, einem relativ generalisierten Datensatz, der insgesamt 5479 einzelne Waldparzellen enthält. Aus dem anfänglichen Flickr-Datensatz, der mehr als zwei Millionen Einzelstandorte enthielt, blieben uns 145000, die von 5400 Einzelnutzern über den Zeitraum von 13 Jahre generiert wurden. Wir gingen davon aus, dass eine Person eine Waldparzelle normalerweise nicht mehr als einmal pro Tag besucht, und berechneten die Besuche einzelner Benutzer auf Waldparzellen pro Tag als unsere grundlegende Analyseeinheit.

Wir führten Analysen auf zwei räumlichen und zwei zeitlichen Skalen durch. Die grundlegende räumliche Skala untersuchte einzelne Waldparzellen und hat die Anzahl der einmaligen Besuche pro Tag, aggregiert über einzelne Jahre, gezählt. Da sich, wie in Abbildung 2 dargestellt, die absoluten Zählungen der beigetragenen Flickr-Bilder (und der Benutzer*innen) im Laufe der Zeit variiert haben, normalisierten wir die Zählungen der Waldnutzer*innen nach der Gesamtzahl der in der Schweiz in einem bestimmten Jahr aktiven einmaligen Flickr-Nutzer*innen. Ausserdem werden grössere Waldparzellen aufgrund ihrer Fläche wahrscheinlich eine grössere Anzahl von Besuchen aufweisen. Wir haben deshalb auch die Zählungen nach Waldfläche normalisiert.

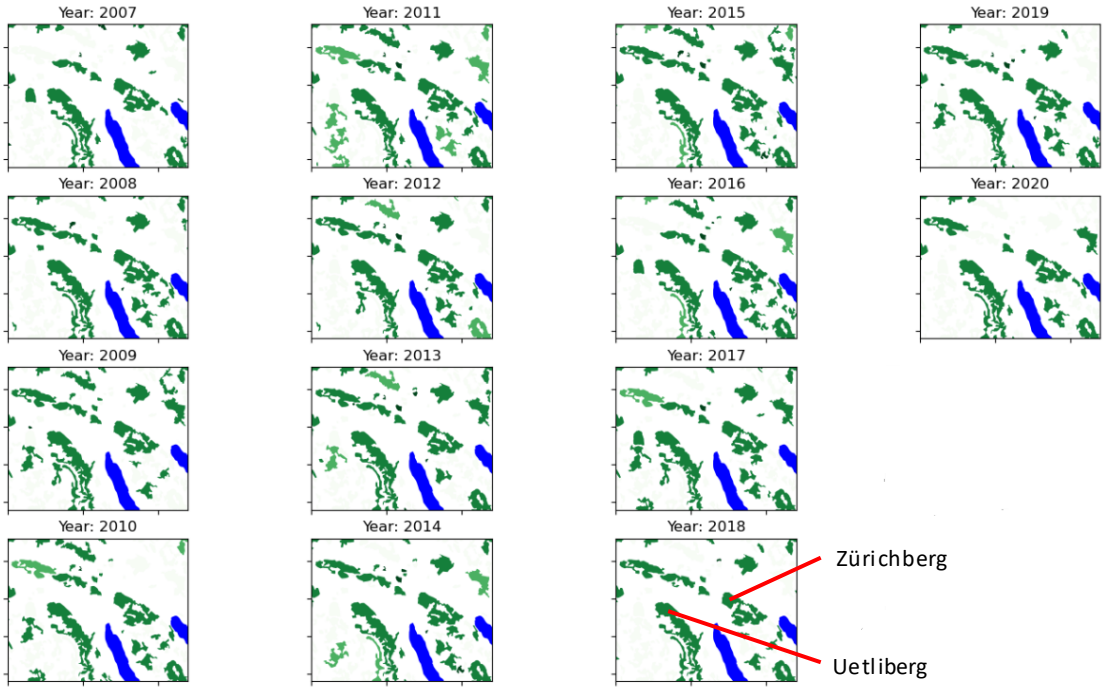


Abbildung 7: Täglich normalisierte Zählungen von Flickr-Nutzer*innenbesuchen in Wäldern, aggregiert jährlich für die Region Zürich. Dunkle Farben zeigen eine höhere Anzahl von Besuchen an.

Abbildung 7 zeigt die Ergebnisse dieses Ansatzes für einzelne Parzellen im Raum Zürich. Man beachte die anhaltend hohen relativen Zählungen für wichtige Erholungswaldgebiete wie Uetliberg und Zürichberg und wie andere Waldflächen im Laufe der Zeit erscheinen und verschwinden. Im Allgemeinen waren die resultierenden Datenmengen pro Waldparzelle jedoch gering. Basierend auf dem Gesamtdatensatz, der mit Wäldern assoziiert ist, wurden pro Parzelle und Jahr durchschnittlich etwa zwei Bilder pro Parzelle gezählt. Das bedeutet, dass die Untersuchung einzelner Waldparzellen mit diesen Daten wahrscheinlich nicht effektiv sein wird, insbesondere angesichts der Veränderungen in der Nutzung der Plattform im Laufe der Zeit.

Wir aggregierten deshalb Waldparzellen nach biogeografischen Regionen der Schweiz (Gonseth et al. 2001) und berechneten für diese Regionen einmalige Parzellenbesuche für die meteorologischen Jahreszeiten (Frühling, Sommer, Herbst und Winter) zwischen 2007 und 2020. Abbildung 8 zeigt Histogramme der absoluten Zählungen für jede Region, während in Abbildung 9 normalisierte Werte für Waldbesuche dargestellt sind. Diese Darstellungen zeigen deutlich die Bedeutung des Mittellandwaldes für Erholungszwecke. Die zweitbeliebteste Region ist die Alpennordflanke, wahrscheinlich als Ergebnis von Fotografien, die auf Skipisten und im Sommer und Herbst auf Wanderwegen aufgenommen wurden. Die normalisierten Zählungen zeigen insbesondere niedrige relative Zählungen für den Jura (die sich höchstwahrscheinlich aus der grossen Waldfläche in Abhängigkeit von der Gesamtfläche der Region und einer relativ kleinen Zahl von Nutzern ergeben, die zu diesen Flächen beitragen). Deutliche saisonale Schwankungen sind ebenfalls sichtbar, vor allem die viel geringere Popularität des Waldes in allen biogeographischen Regionen im Winter. In unserer dritten und letzten Fallstudie haben wir untersucht, welche Faktoren Menschen in verschiedenen Jahreszeiten in Waldgebiete locken oder sie davon fernhalten. Dazu müssen wir über die Analyse auf der Grundlage von Nutzern, Zeit und Ort hinausgehen und die inhaltliche Semantik untersuchen, wie sie in den mit den Bildern verbundenen Metadaten erfasst wird.

Abbildung 8: Histogramme der absoluten Zählungen von saisonal einmaligen Waldparzellenbesuchen für die biogeografischen Regionen der Schweiz

Abbildung 9: Histogramme von normalisierten saisonalen einmaligen Waldbesuchen auf Waldparzellen für Schweizer biogeografische Regionen

4.3 Fallstudie drei: Semantik der Erholungswaldnutzung

In dieser Fallstudie untersuchten wir, wie Menschen Wald, mit Hilfe der den Bildern zugewiesenen Tags, beschreiben. Wir untersuchten drei Forschungsfragen:

- Was verraten uns die zur Beschreibung des Waldes verwendeten Tags über das Verhalten der Besucher*innen?
- Können wir gemeinsam auftretende Tags verwenden, um bestimmte Verhaltensweisen in Raum und Zeit abzubilden?
- Gibt es Unterschiede in Bezug auf die Sprache?

Um diese Fragen zu klären, haben wir zunächst den Originaldatensatz gefiltert, um Massen-Uploads zu entfernen. Dabei wurden jeweils Datenpunkte mit einer identischen Tag-Liste, sowie einer identischen Benutzer*innen ID und einem identischen Datum entfernt. Auf diese Weise wurde der Datensatz auf etwa 900'000 Einträge reduziert.

Unser Ansatz basierte auf der einfachen Annahme, dass Bildmetadaten, welche das Schlüsselwort `Forest` oder sprachlich verwandte Terme im Französischen oder Deutschen (z.B. `forêt` oder `Wald`) enthalten, mit Waldbesuchen in Verbindung gebracht werden können. Um zu untersuchen, wie Wald im Datensatz wahrgenommen wurde, untersuchten wir Begriffe, die zusammen in Taglisten vorkommen. Anstelle von Rohfrequenzen (die typischerweise von sehr häufigen Begriffen dominiert werden) berechneten wir Dice-Scores, die sehr exklusive, aber nicht seltene Kombinationen von Tags darstellen. Für einen Begriff (`Forest`) extrahierten wir die hundert nach der Dice-Scores am häufigsten vorkommenden Tags und annotierten diese als Elemente (konkrete Objekte, die wahrscheinlich in einem Bild vorkommen), Wetter, Aktivitäten (z.B. Wandern oder Radfahren) und Eigenschaften (Eigenschaften von Bildern wie z.B. schön) und identifizierten wahrscheinliche

Duplikate. Abbildung 10 zeigt die 47 einzigartigen Kognaten, die als mit `Forest` zusammen vorkommend identifiziert wurden.

Abbildung 10: 47 einzigartige, mit `Wald` assoziierte Terme

Diese Begriffe bestehen aus 21 Objekten (z.B. leaf, tree, mushroom, trail), 7 Wetterelementen (z.B. fog, sun, cloud), 15 Qualitäten (z.B. landscape, nature, beautiful) und 4 Aktivitäten (hiking, walking, riding, travel). In Abbildung 11 stellen wir die Verwendung ausschliesslich französischer oder deutscher Wald-Kognate (`Wald` oder `Forêt`) in Bild-Metadaten dar und stellen fest, dass diese stark die Regionen widerspiegeln, in denen in der Schweiz französisch und deutsch gesprochen wird . Somit ist es höchstwahrscheinlich, dass diese Beiträge von Schweizerinnen und Schweizern geleistet wurden und die Naherholungsnutzung des Waldes in der Schweiz widerspiegeln .

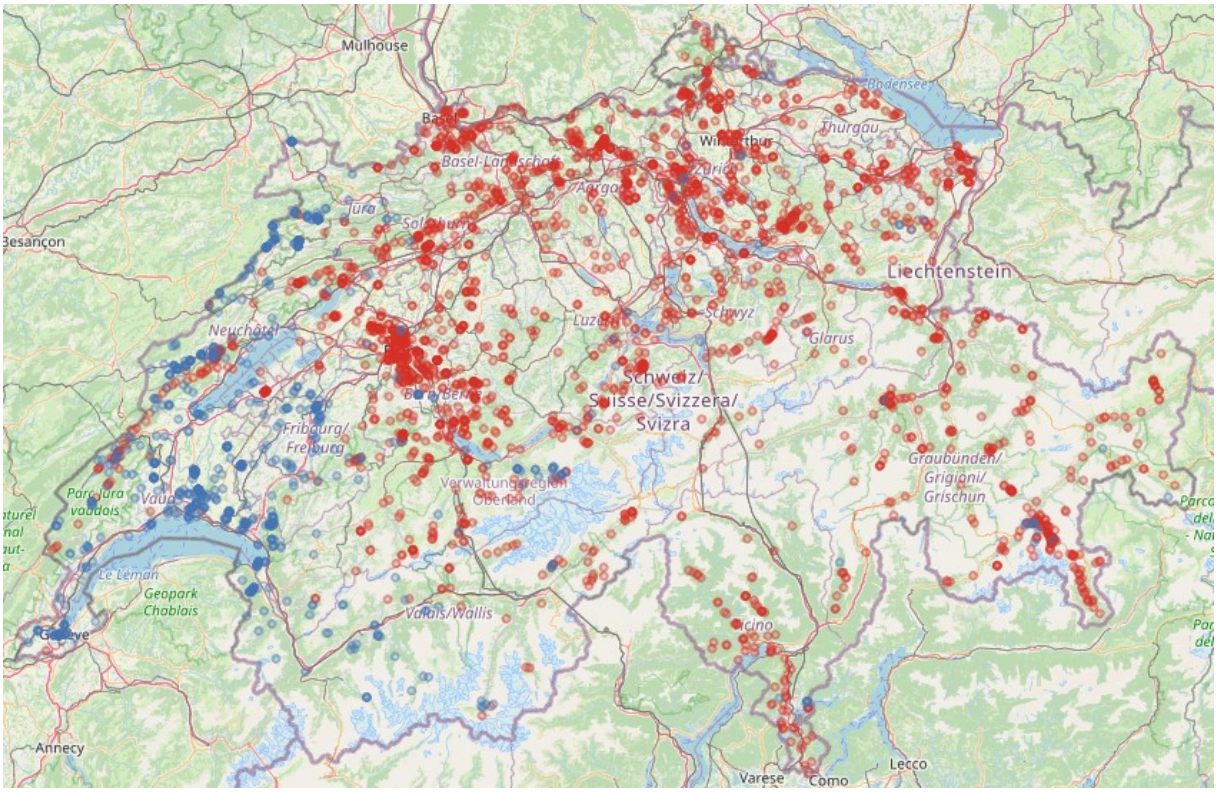


Abbildung 11: Position von Bildern, die ausschliesslich mit `Wald` (rot) oder `Forêt` (blau) getaggt sind. Hintergrund Kartierung © OpenStreetMap contributors

Abbildung 12 zeigt die am höchsten eingestuften Terme für die Begriffe `forest`, `Wald` und `forêt` für die Jahreszeiten Frühling und Herbst. Alle sechs Wortwolken wurden mit den 50 am höchsten bewerteten Termen erstellt, wobei Terme mit niedrigeren Punktzahlen in kleinerer Schriftgrösse aufgetragen wurden. Wortwolken, in denen mehr Begriffe gut sichtbar sind, lassen somit auf eine gleichmässigere Verteilung der Begriffsverwendung schliessen. Wir stellen hier zwei Punkte fest. Erstens, dass der im Herbst verwendete Wortschatz im Allgemeinen reicher zu sein scheint und zweitens, dass die im Französischen verwendeten Begriffe vielfältiger erscheinen. Wir stellen zudem fest, dass `forest` in der Sammlung am häufigsten vorkommt (9089 Fälle), gefolgt von `Wald` (5621) und schliesslich `forêt` (2245).

Es sticht ins Auge , dass «Mushroom» und ähnliche Terme besonders häufig mit den Begriffen `forest`, `Wald` und `forêt` vorkommen. In Abbildung 12 extrahieren wir deshalb alle Bildstandorte, die mit verwandten Begriffen aus dem Englischen, Deutschen und Französischen in Zusammenhang mit Pilzen stehen. Die identifizierten Begriffe lauten: `mushroom`, `pilz`, `champignon`, `fungi` and `fungus`. Von Interesse sind hier die Bild-Cluster, die sich auf Pilze beziehen und sich um grössere Schweizer Städte (z.B. Zürich, Basel, Bern und Genf) gruppieren. Durch die Untersuchung der zugehörigen Bilder wäre es möglich, weiter zu untersuchen, ob diese Bilder in erster Linie mit Ästhetik und Individuen assoziiert werden, die sich gerne Pilze im Wald anschauen, oder ob sie tatsächlich auf das Sammeln von Pilzen in periurbanen Wäldern hinweisen.

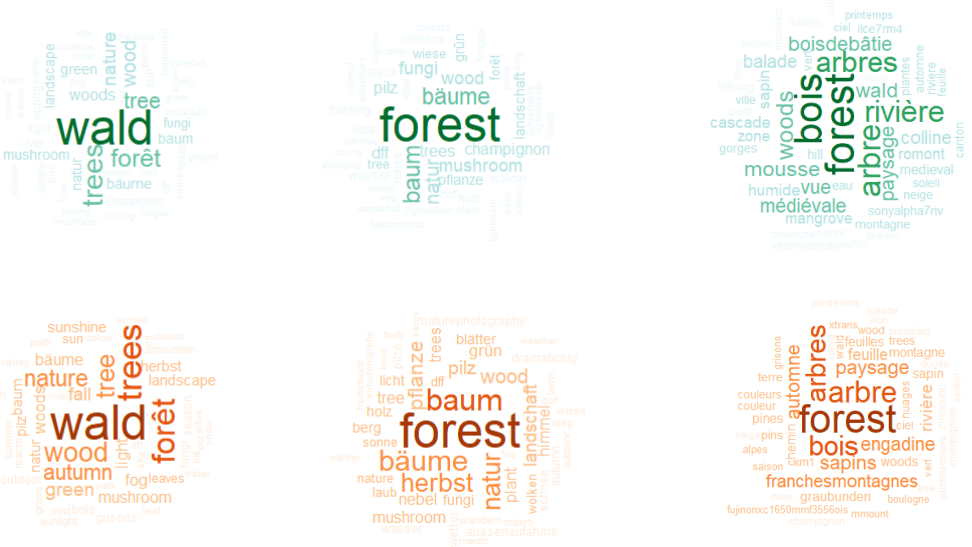


Abbildung 12: Höchstplatzierte Terme für die Begriffe `forest`, `Wald` und `forêt`

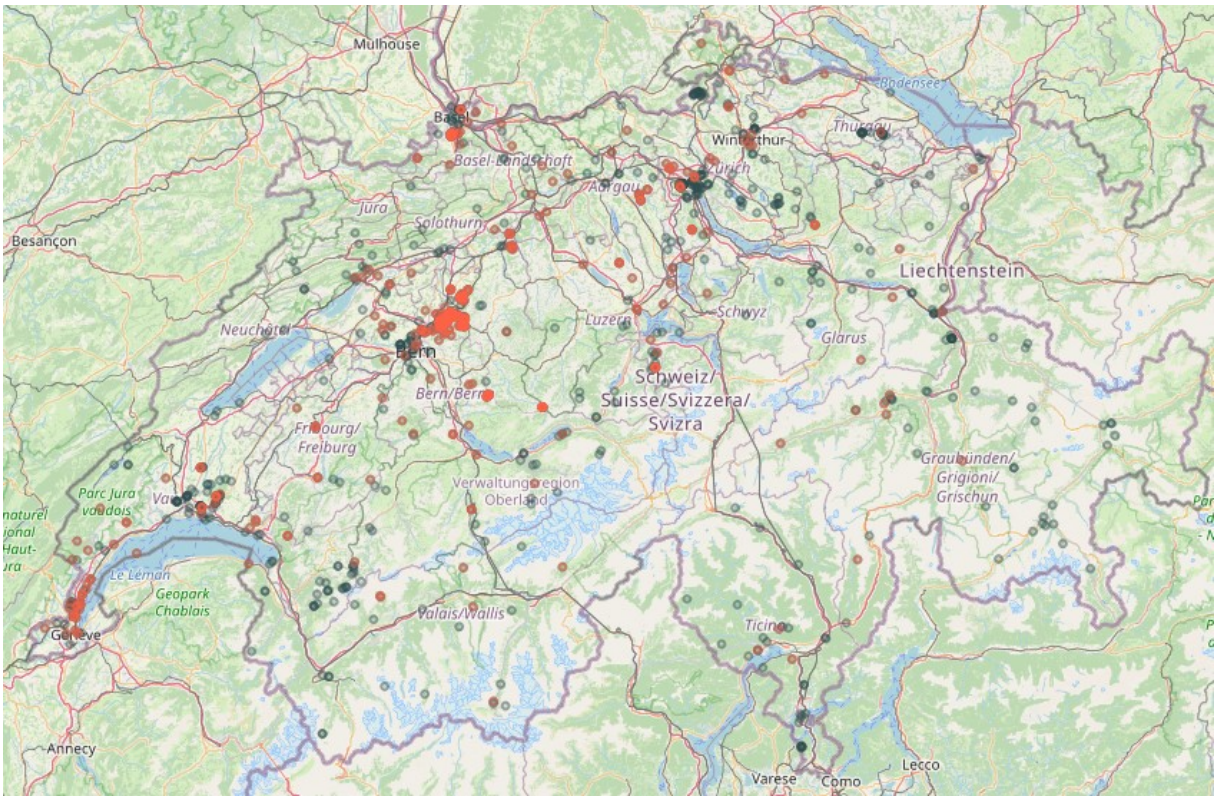


Abbildung 11: Lage der Bilder mit gleichzeitigem Vorkommen von Kognaten von `Pilz` und `Wald` (rot) und Verwandte des `Pilzes` (blau) Hintergrundkartierung © OpenStreetMap contributors

5. Abschliessende Bemerkungen

Unser ursprünglicher Auftrag für diese Untersuchung war, die Analysen von Wartmann et al. (2018) zu wiederholen. Aufgrund signifikanter Veränderungen beim Zugang zu zwei der Plattformen (Twitter und Instagram) passten wir unsere Ergebnisse jedoch an, um uns auf eine einzige Quelle (Flickr) zu konzentrieren. Diese Veränderungen zeigen einmal mehr, dass die Erstellung von Langzeitindikatoren auf der Grundlage von Social Media-Daten eine Strategie mit hohem Risiko ist, und zwar sowohl aufgrund von Veränderungen beim Zugang als auch, wie in Abbildung 2 dargestellt, aufgrund von Veränderungen in der Popularität von Social Media-Plattformen.

Nichtsdestotrotz bieten Social Media-Daten (wie zum Beispiel die verwendeten Flickr-Daten) sehr reichhaltiges Material für die Analyse und den Vergleich mit bestehenden Ansätzen. Wie wir in der ersten Fallstudie gezeigt haben, können diese Daten zur Untersuchung des Freizeitpotenzials verwendet werden und, was ebenso wichtig ist, auf potenzielle Probleme mit bestehenden Modellen hinweisen.

In unserer zweiten Fallstudie extrahierten wir einzigartige tägliche Besuche von Waldparzellen in der Schweiz. Diese Studie demonstrierte die methodische Bedeutung der Normalisierung der Daten sowohl für Veränderungen in der Nutzung von sozialen Medien als auch für die untersuchte Waldfläche. Auf diese Weise, und durch Aggregation über die biogeografischen Regionen der Schweiz hinweg, konnten wir die räumliche und zeitliche Variation in der Nutzung der Schweizer Wälder und, vor allem, die Bedeutung der Mittellandwälder klar aufzeigen.

Durch die Analyse der Semantik anhand der Tags, die mit Waldbildern verbunden sind, konnten wir mehr darüber erfahren, warum Personen die Schweizer Wälder besuchen. Es überrascht nicht, dass Fotografen*innen in erster Linie visuelle Elemente der Waldszene wahrnehmen und beschreiben, und wir stellten die Bedeutung wetterbezogener Tags fest, insbesondere im Herbst, und zeigten, wie räumliche Muster, die mit bestimmten Elementen zusammenhängen, erforscht werden können. Die Verteilung der Wald-Kognate deutet stark darauf hin, dass der von uns untersuchte Inhalt von lokalen Benutzern hochgeladen wurde, und weist auf seinen Wert bei der weiteren Untersuchung der Nutzung der Schweizer Wälder hin.

Auf der Grundlage dieser Studie geben wir dem BAFU drei Empfehlungen:

- Die Erstellung von Langzeitindikatoren unter Verwendung von Social Media-Daten sollten nicht zum Ziel haben, andere Ansätze zu ersetzen, insbesondere aufgrund des instabilen und unvorhersehbaren Zugangs zu externen Datenquellen.
- Die Entwicklung, Unterstützung und Verwaltung von bürgerwissenschaftlichen, partizipativen Projekten über lange Zeiträume ist ein möglicher Ansatz, um Daten ähnlich wie bei Flickr zu sammeln, und ein Ansatz, bei dem das BAFU datenschutzbezogene Fragen besser kontrollieren und die Nutzer*innen motivieren könnte.
- Social Media-Daten sind sehr reichhaltig. Durch die Kombination verschiedener Dimensionen (wie in den hier vorgestellten Fallstudien demonstriert) können sie dazu verwendet werden, Hypothesen zu entwickeln und zu testen. Sie dienen auch der Erforschung von bestehende Modellen - solche Studien sollten durchgeführt werden, um bestehende Arbeiten zu ergänzen und potenzielle neue Wege für die Forschung vorzuschlagen.

Literatur

Chesnokova, O., Taylor, J. E., Gregory, I. N., & Purves, R. S. (2019). Hearing the silence: finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the English Lake District. *International Journal of Geographical Information Science*, *33*(12), 2430-2454.

Gonseth, Y., Wohlgemuth, T., Sansonnens, B., & Buttler, A. (2001). Die biogeographischen Regionen der Schweiz-Les régions biogéographiques de la Suisse. *Bern/Berne: BUWAL/OFEFP*.

Haklay, M. M. (2016). Why is participation inequality important? *European Handbook of Crowdsourced Geographic Information*, 35.

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 2010(1), 21-48.

Hu, Y., & Wang, R. Q. (2020). Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour*, 1-3.

Llewellyn, C., & Cram, L. (2016). Brexit? analyzing opinion on the UK-EU referendum within Twitter. In *Tenth International AAAI Conference on Web and Social Media*.

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2020). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-150, https://CRAN.R-project.org/package=nlme.

Purves, R. S., & Derungs, C. (2015). From space to place: Place-based explorations of text. *International Journal of Humanities and Arts Computing*, *9*(1), 74-94.

Purves, R. S., & Mackaness, W. A. (2016). A methodological toolbox for exploring collections of textually annotated georeferenced photographs. *European Handbook of Crowdsourced Geographic Information*, 145.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H. and Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, *233*, pp.298-315., 46, 933–945.

Wartmann, F. M., Acheson, E., & Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: an interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8), 1572-1592.

Wartmann, F.; Bär, M.; Purves, R.; Hunziker, M. (2018): Das Potential von Daten aus sozialen Medien für die Erforschung der Walderholung. Pilotprojekt als ergänzendes Testmodul zum Projekt «WaMos meets LFI» (WML). Im Auftrag des Bundesamtes für Umwelt (BAFU). Interner Bericht an den Auftraggeber. Birmensdorf, Eidg. Forschungsanstalt für Wald, Schnee und Landschaft, WSL.