Mapping the Crowd – Zur Rolle der Mapper bei der Qualitätsanalyse von OpenStreetMap

Christopher BARRON, Pascal NEIS, Alexander ZIPF

Zusammenfassung

Web 2.0 Projekte wie Wikipedia oder *OpenStreetMap* (OSM) gewinnen an Beliebtheit und die Nutzergruppen wachsen stetig. Nachdem sich bisherige Untersuchungen insbesondere auf die Qualität der gesammelten Informationen bezogen, tritt die Community selbst immer mehr in den Vordergrund wissenschaftlicher Untersuchungen. Dabei werden beispielsweise bestimmte Charakteristika des Verhaltens der Beitragenden untersucht, um daraus z.B. indirekt Aussagen über die Qualität der Daten ableiten zu können. Um derartige Untersuchungen zu ermöglichen, werden in diesem Artikel unterschiedliche Verfahren zur Ermittlung der Quantität, Aktivität und Art der Beiträge von Mitgliedern präsentiert. Die Ergebnisse können helfen die Qualität von OSM Daten einzuschätzen. Die vorgestellten Verfahren verwenden dabei ein OSM-Full-History-Dump, wodurch intrinsische Qualitätsmaße abgeleitet werden können. Hierzu wird der Stand der Wissenschaft und die Umsetzung der entsprechenden Analyseverfahren erläutert und diskutiert. Exemplarisch werden Ergebnisse bezogen auf die Stadt Heidelberg vorgestellt.

1 Einleitung

Das OSM Projekt hat seit Anfang des Jahres 2013 über eine Million registrierte Mitglieder (OGD 2013). Der aktuelle Datenbestand (April, 2013) umfasst weltweit über 1,8 Milliarden Punkte und insgesamt knapp 180 Million Linien und Polygone. Bisherige Qualitätsuntersuchungen über OSM konzentrierten sich anfänglich stark auf das Straßennetzwerk. Dabei wurde dieses für unterschiedliche Länder mit verschiedenen Referenzdatensätzen verglichen (z. B. Ordnance Survey Meridian 2, TeleAtlas-MultiMet/TomTom, Navteq, HMGS). Mit der Zeit richtete sich der Fokus auch auf andere Objekte des OSM Projektes. Verschiedentlich wird in der Literatur der letzten Jahren angemerkt, dass für OSM Qualitätsuntersuchungen oftmals keine Referenzdaten zur Verfügung stehen und alternative Wege zur Bewertung der Daten gefunden werden müssten (MOONEY ET AL. 2010, MOONEY & CORCORAN 2011). Eine Alternative besteht dabei in der Analyse der OSM-Datenhistorie. In diesem Zusammenhang führen VAN EXEL ET AL. (2010) erstmals den Begriff der *Crowd Quality* (CQ) ein. In diesem zweidimensionalen Ansatz steht nicht mehr nur alleine die Qualität der (Geo-) Daten (*Feature Quality*) im Vordergrund, sondern erstmals auch die Qualität der Mitwirkenden (*User Quality*) und das Abhängigkeitsverhältnis der beiden zueinander (*Interdependency*). Diese Informationen können aus der Historie eines VGI-Datensatzes abgeleitet werden. Der Ansatz eignet sich daher insbesondere für sog. „crowdsourced datasets“, also für Datensätze, die kollaborativ von einer *crowd* (Menge aller Mitwirkenden) erschaffen werden. Im Gegensatz zu institutionellen Geodatensammlungen

müssen die Beteiligten kein Vorwissen oder jegliche Art an Qualifikation mitbringen, um zum Projekt beizutragen. Die *User Quality* wird von drei wesentlichen Faktoren bestimmt:

- „*Local Knowledge*“: Lokales Wissen bzw. die Vertrautheit der Mitwirkenden mit einer Gegend,
- „*Experience*“: Erfahrung des Mitwirkenden in dem Projekt,
- „*Recognition*“: Anerkennung des Mitwirkenden innerhalb des Projekts.

Der Beitrag gliedert sich wie folgt: Abschnitt 2 diskutiert den aktuellen Stand der Forschung. Abschnitt 3 behandelt die Untersuchungen zu den Aktivitäten und Verhalten der Mitwirkenden. Dabei wird insbesondere auf die Aspekte der Entwicklung in der Anzahl der Mitwirkenden, deren Aktivitäten und deren ungleichen Verteilung der Beiträge eingegangen. Eine Diskussion und ein Ausblick schließen den Beitrag ab.

2 Stand der Forschung - Analyse des Mitwirkenden

Wie bereits erwähnt gibt es zahlreiche wissenschaftliche Arbeiten, die sich mit dem Vergleich der bei OSM gesammelten Geodaten zu einem Referenzdatensatz beschäftigen. Um nur ein paar beispielsweise für das Straßennetzwerk zu nennen sind dies HAKLAY (2008) für England, ZIELSTRA & ZIPF (2010), NEIS ET AL. (2012) für Deutschland oder GIRRES & TOUYA (2010) für Frankreich. Nicht immer ist für ein Untersuchungsgebiet jedoch auch ein Referenzdatensatz für einen Qualitätsvergleich verfügbar. Nach CIEPLUCH ET AL. (2011) kann zwischen einfachen und fortgeschrittenen Qualitätsindikatoren für OSM unterschieden werden. Erste sind z. B. einfache Längenvergleiche zwischen zwei OSM Datensätzen, letztere u. a. regionale Editierschwerpunkte oder Mitglieder-Profile, die aus der Datenhistorie abgeleitet werden können. Auf die komplette Datenhistorie eines Objektes konnte bis Ende des Jahres 2009 nur über die Web-API des OSM Projektes zugegriffen werden. Dadurch waren großflächige Analysen der vollständigen Historie der OSM Objekte nicht möglich. Seit Ende des Jahres 2009 gibt es in ca. viertel-jährigen Intervallen jeweils einen vollständigen Dump der OSM Datenbank mit der kompletten Historie aller Objekte. In den folgenden Jahren gab es kaum Programme, die in der Lage waren, diese Daten zu verarbeiten. Inzwischen gibt es jedoch mehre Anwendungen zum Verarbeiten oder Analysieren dieser Daten. MOONEY & CORCORAN (2011a) beschäftigen sich vor allem mit der Historie von sog. „*heavily edited objects*“ (HEO). Dies sind nach ihrer Definition Objekte in OSM mit mehr als 15 Versionen. Da bei diesen Objekten davon ausgegangen werden kann, dass sie in einem hohen Maß editiert und bearbeitet wurden, sind diese für Analysen ihrer Meinung nach am besten geeignet. Für erste Untersuchungen analysieren und extrahieren sie die vollständige Historie von 25.000 dieser HEOs aus mehreren Ländern. Dabei betrachten sie vor allem die Verteilung der „*name*“ und „*highway*“-Tags näher. In einer weiteren Untersuchung dehnen die Autoren ihr Untersuchungsgebiet auf ganz Großbritannien aus und betrachten in verschiedenen Analysen die Historie der HEOs. Zu den wichtigsten Tags zählen laut der Autoren: „*Amenity*“, „*Highway*“, „*Landuse*“, „*Natural*“ und „*Waterway*“. Gegenstand ihrer Analysen sind neben einer Gegenüberstellung von Bearbeitungen und der Summe an Mitwirkenden im Datensatz u.a. die durchschnittliche Anzahl an Beitragenden an einem einzelnen Objekt, verschiedene Zeichenkettenalgorithmen um Ähnlichkeiten zwischen

Tags zu identifizieren und Änderungen in der Geometrievalidität von aufeinanderfolgenden OSM-Objekten (MOONEY & CORCORAN 2012.). NEIS & ZIPF (2012) untersuchen erstmals die Anzahl, Herkunft und Aktivität der Mitglieder des gesamten OSM Projektes. Sie analysieren u. a. das Verhältnis von registrierten und aktiven Mitgliedern. Von allen registrierten Mitwirkenden haben 38 % jemals mindestens einen Edit vorgenommen und lediglich 5 % mehr als 1.000 Nodes erstellt und damit in einer, wie die Autoren sinngemäß konstatieren, „produktiven Weise zum Projekt beigetragen". Darüber hinaus untersuchen sie die Arbeitsgebiete und Zeitfenster der Mitwirkenden und stellen ihre Ergebnisse für verschiedene Länder vor, wobei letztlich mehr als 2/3 (72 %) aller registrierten User aus Europa stammen.

Weitere Untersuchungen gehen inzwischen verstärkt auf die Mitwirkenden ein, die hinter den Daten eines VGI-Projekts stehen. So stellen beispielsweise REHRL ET AL. (2012) ein grundsätzliches und allgemeines Modell vor, um Beitragsmuster in VGI-Projekten zu analysieren. MOONEY & CORCORAN (2012a) und MOONEY & CORCORAN (2012b) erstellen Mitglieder-Profile aus der OSM-Datenhistorie mit dem Ziel, soziale Netzwerkstrukturen und Interaktionen zwischen den Beitragenden zu konstruieren. Hierfür visualisieren sie u. a. die Editierinteraktionen der 250 am stärksten zum OSM Gebiet von London beitragenden Mitgliedern. Dabei kommen sie jedoch zu dem Ergebnis, dass ein eindeutiges soziales Netzwerk unter den Mitwirkenden nicht ausgemacht werden kann.

Eine von NAPOLITANO & MOONEY (2012) entwickelte Anwendung verfolgt den bereits erwähnten Ansatz der *Crowd Quality* weiter. Sie ermöglicht es, in einem OSM Datensatz Gegenden zu identifizieren, in der sog. „*High Quality User*" neue Objekte hinzufügen oder editieren. Hierfür werden die drei wesentlichen Faktoren der *User Quality* operationalisiert, um anschließend die Mitwirkenden zu identifizieren, deren Beiträge als von besonders hoher Qualität angesehen werden können. Außerdem können mit der Anwendung sog. „*pet locations*" der einzelnen Mitwirkenden im Gebiet berechnet werden. Der Begriff „*pet*" wurde von LIEBERMANN & LIN (2009) bei Untersuchungen der Editier-Historien von Wikipedia-Artikeln verwendet. Dabei stellen sie fest, dass Artikel, welche die Wikipedia-User bearbeiten oder erstellen, oft einen engen geographischen Bezug zu ihrem Wohn- oder Geburtsort haben und von ihnen über eine längere Zeit hinweg weiter bearbeitet werden. NAPOLITANO & MOONEY (2012) übertragen den Begriff der „*pet location*" und bezeichnen damit ein geographisches Gebiet in OSM, das von einem Mitwirkenden besonders gut bearbeitet und betreut wird. NEIS ET AL. (2012a) untersuchen das Problem des Vandalismus in OSM. Dabei zeigen sie in ihrer Studie, dass rund ein ¾ aller erkannten Vandalismusfälle des Projekts von neuen Mitgliedern ausgeht.

3 Untersuchung der Aktivität & Verhalten der Mitwirkenden

Die Qualität der OSM-Daten kann bis zu einem bestimmten Maß über die Analyse der Daten an sich ermittelt werden. Bei VGI-Projekten nehmen insbesondere die Personen, die die Daten erstellen, eine wichtige Rolle im Hinblick auf die Qualität der Daten ein. In den meisten Online-Communities kann ein Phänomen festgestellt werden, das sich mit „*Participation Inequality*" (dt.: „Ungleichheit in der Partizipation") beschreiben lässt und das bei der Interpretation der Ergebnisse stets berücksichtigt werden muss. „*Participation Inequality*" besagt, dass die Teilnahme an den meisten Online-Communities einem 90-9-1 Schema folgt: 90 % aller Nutzer konsumieren ohne zu dem Projekt beizutragen, 9 % tragen gele-

gentlich dazu bei und lediglich 1 % der Nutzer trägt in einem hohen Maße zum Projekt bei (NIELSEN 2006). Dass dieses Phänomen auch für das OSM Projekt zutrifft zeigen, wie bereits erwähnt, NEIS & ZIPF (2012). In den folgenden Unterkapiteln geht es speziell um die Mitglieder, die gelegentlich oder in einem hohen Maße zu OSM beitragen. Es werden Methoden und Analysen vorgestellt, die Informationen zu diesen hinter den Daten stehenden Mitgliedern liefern.

3.1 Anzahl der Mitwirkenden

Die Zahl der Mitwirkenden in einem OSM Gebiet gibt Auskunft darüber, wie viele Mitwirkende für die erfassten und bearbeiteten Daten verantwortlich sind. Da in zahlreichen Untersuchungen festgestellt wurde, dass eine große Anzahl an aktiven Mitwirkenden zu einem qualitativ besseren Datensatz führt, gibt diese Analyse auch Auskunft über die mögliche Qualität des Datensatzes (MOONEY & CORCORAN 2012b). Eine Möglichkeit dies zu ermitteln besteht in der Untersuchung der historischen Entwicklung der Gesamtzahl aller Mitwirkenden. Aufgrund der erwähnten „*Participation Inequality*“ sollte zudem die Zahl der pro Monat aktiven Mitwirkenden betrachtet werden. Aus diesen beiden Parametern können Rückschlüsse auf die (monatliche) Aktivität der Mitwirkenden gezogen werden.

Die Gesamtzahl aller Mitwirkenden zu einem Zeitpunkt wird dadurch bestimmt, dass die Zahl der Mitwirkenden eines jeden Monats, die mindestens einen Node erstellt oder bearbeitet haben, bis zu diesem Zeitpunkt aggregiert werden. Die Summe dieser wird in einem Histogramm über die vollständige Historie des Datensatzes dargestellt. Die Anzahl der monatlich aktiven Mitwirkenden wird dadurch errechnet, dass die Summe aller Mitwirkenden pro Monat ermittelt wird. Im Gegensatz zu erstgenannter Analyse werden die Ergebnisse hierbei jedoch nicht aggregiert.

Grundsätzlich ist es für die Qualität von OSM Daten wichtig, dass möglichst viele Mitglieder in der Erfassung und Bearbeitung involviert sind. Eine hohe Anzahl an Mitwirkenden hat zur Folge, dass mehr Menschen für die Daten verantwortlich sind und dadurch möglicherweise mehr Daten erfasst bzw. mehr Fehler identifiziert und anschließend korrigiert werden. Zu Beginn der Entwicklung von OSM in einer Region kann angenommen werden, dass die Gesamtzahl der Mitwirkenden, die jemals einen Node beigetragen haben, verhältnismäßig gering ist. Der erste deutlichere Ausschlag deutet auf den Beginn einer größeren Aktivität und, zumindest auf regionaler Ebene, auf die Entstehung einer Community in einem Gebiet hin. Im Idealfall nimmt die Zahl der Mitwirkenden in der Folge kontinuierlich zu. Eine Stagnation ist jedoch nicht gleichzusetzen mit einem Stillstand der Aktivität oder dem generellen Innehalten in der Entwicklung eines OSM-Gebiets. Sie besagt lediglich, dass keine neuen Mitwirkenden hinzugekommen sind.

Um einen differenzierten Blick auf die monatliche Situation zu bekommen, sind die Ergebnisse weiterer Analysen notwendig. Die Entwicklung der Anzahl der pro Monat tatsächlich aktiven Mitwirkenden ermöglicht weitergehende Analysen. Bei einer Gegenüberstellung der beiden Histogramme kann ermittelt werden, wie viele neue Mitglieder die Daten erfasst oder bearbeitet haben. Je höher diese Anzahl der Mitwirkenden im Durchschnitt ist, desto größer ist die Heterogenität und desto höher ist möglicherweise auch die Qualität der Daten in einem Gebiet. Abbildung 1a und Abbildung 1b visualisieren die Ergebnisse der beiden Berechnungen für das exemplarische Untersuchungsgebiet von Heidelberg.

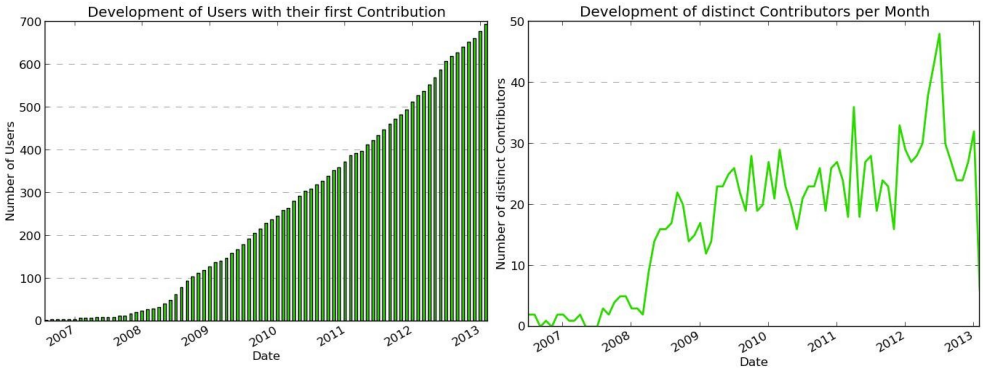

Abb. 1: a) Monatliche Entwicklung der aggregierten Anzahl an Mitwirkenden in Heidelberg b) Monatliche Entwicklung der Anzahl an eindeutigen Mitwirkenden

In beiden Diagrammen ist erkennbar, dass die Zahl der Mitwirkenden ab Anfang 2008 deutlich zunimmt. Ab Juni 2010 pendelt sich die Zahl der monatlich aktiven Mitwirkenden bei durchschnittlich 25 (Abbildung 1b) ein. Die aggregierte Zahl derjeniger, die jemals einen Node erstellt haben (Abbildung 1a), nimmt jedoch kontinuierlich weiter zu. Dies veranschaulicht, warum die beiden Parameter getrennt voneinander betrachtet werden müssen und nicht aus der Menge auf die tatsächlich aktiven Mitwirkenden geschlossen werden kann.

3.2 Aktivität der Mitwirkenden

NEIS & ZIPF (2012) ermitteln in ihrer Untersuchung des weltweiten OSM Datensatzes die Projektaktivität aller Mitglieder. Diese teilen sie anhand der Anzahl ihrer erstellten Nodes in vier Nutzergruppen ein: „*Senior Mappers*“, „*Junior Mappers*“, „*Nonrecurring Mappers*“ und „*Mapper ohne Node-Beitrag*“. Mit 62 % aller registrierten Mitwirkenden stellen die Mapper ohne einen Beitrag die mit Abstand größte Gruppe dar, gefolgt von den „*Nonrecurring Mappers*“ (19 %), den „*Junior Mappers*“ (14 %) und den „*Senior Mappers*“ (5 %). Auf Grundlage dieser Kategorisierung kann ein Gebiet in OSM anhand seiner Mitwirkenden klassifiziert werden.

Das hier gezeigte Vorgehen orientiert sich an NEIS & ZIPF (2012). Insgesamt werden jedoch nur drei Kategorien gebildet, da in der OSM Datenhistorie keine Zahlen zu denjenigen Mitgliedern vorliegen, die sich zwar bei OSM registriert, jedoch nie einen Node erstellt haben. Vier Kategorien sind nur bei weltweiten Untersuchungen möglich, bei denen die Zahl der Mitwirkenden mit Bearbeitungen im Datensatz der Gesamtzahl aller registrierten Mitglieder gegenübergestellt werden kann. Die Ergebnisse werden in einem Kreisdiagramm dargestellt und die drei Kategorien der Analyse wie folgt definiert:

- „*Senior Mappers*“: Mitwirkende mit 1000 und mehr erstellten Nodes,
- „*Junior Mappers*“: Mitwirkende mit mindestens 10 und weniger als 1000 erstellten Nodes,
- „*Nonrecurring Mappers*“: Mitwirkende mit weniger als 10 erstellten Nodes.

Diese Analyse zeigt, wie die Mitwirkenden im zugrundeliegenden OSM Gebiet kategorisiert werden können. Je mehr Mitwirkende in den Kategorien mit vielen Node-Erstellungen anzusiedeln sind, desto mehr aktiv Beitragende gibt es in einem Gebiet. Die Gruppe der „*Senior Mappers*“ steht aufgrund der „*Participation Inequality*“ in einem ungleichen Verhältnis zu den anderen beiden. Je geringer diese Ungleichheit ausfällt, desto besser für die Quantität und u. U. auch die Qualität der Daten. Es muss jedoch beachtet werden, dass bei der Analyse eines Gebiets in OSM nicht argumentiert werden kann, dass eine hohe Anzahl von „*Junior Mappers*“ oder „*Nonrecurring Mappers*“ gleichzeitig für Mitwirkende stehen, die weniger Beiträge zu OSM insgesamt erbracht haben. Die Aussagen dieser Analyse beziehen sich nur auf das untersuchte Gebiet und sind in hohem Maße auch von dessen Größe abhängig. Unter Umständen wird ein Mitglied in dem einen Bereich als „*Nonrecurring Mapper*“ identifiziert und in einer anderen Gegend als „*Senior Mapper*“ eingestuft. Somit darf in dieser Analyse nicht von der Quantität der Beiträge auf die Qualität des Mitwirkenden im Allgemeinen gefolgert werden. Dies wäre lediglich dann möglich, wenn die Analyse für die komplette OSM Datenbank durchgeführt werden würde und zudem angenommen wird, dass Beiträge von Mitgliedern mit einer hohen Anzahl an erstellten Nodes aufgrund einer größeren Projekterfahrung (qualitativ) besser sind als von solchen, die weniger Nodes erfasst haben. Für das exemplarische Untersuchungsgebiet Heidelberg zeigt das Kreisdiagramm in Abbildung 2 die Kategorisierung der Mitwirkenden. Mit 4,9 % fällt darin der Anteil der „*Senior Mappers*“ am geringsten aus.

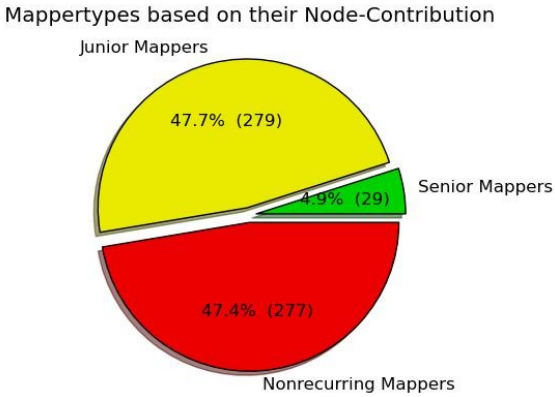

Abb. 2: Kategorisierung der Mitwirkenden [prozentuale (absolute Angaben)] auf Grundlage der Anzahl an erstellten Nodes in Heidelberg

Für weitergehende Untersuchungen wäre es denkbar, in einem ersten Schritt die Kategorisierung der Mitwirkenden auf einer weltweiten Datengrundlage zu berechnen. Dadurch kann jedem jemals aktiv gewordenem Mitwirkenden in OSM eine Kategorie zugewiesen werden. In einem zweiten Schritt kann diese Kategorisierung mit den im jeweiligen Gebiet aktiven Mitwirkenden verknüpft werden. Dies ermöglicht unter Umständen neue Aussagen zur Art der Mitwirkenden in dem jeweiligen Gebiet und damit eventuell zur Qualität der von ihnen erstellten Daten.

3.3 (Un)gleichverteilung der Beiträge

Wie bereits erwähnt kann in der OSM Community eine ungleiche Partizipation der Mitglieder hinsichtlich der Datenerstellung und -bearbeitung ausgemacht werden. Zumeist ist ein kleiner Kreis an Mitwirkenden für einen großen Anteil der Daten verantwortlich, was sich in sog. „*long-tailed distributions*“ äußert (MOONEY & CORCORAN 2012). Die Analyse der potentiellen Ungleichheit in einem Gebiet kann Auskunft über die anteiligen Beiträge der Mitwirkenden geben. Dies geschieht wie im Folgenden dargelegt:

Für jeden Mitwirkenden, der entweder einen Punkt, eine Linie oder ein Polygon in einem ausgewählten OSM Gebiet erstellt hat, wird sein entsprechender prozentualer Anteil an erstellten Objekten berechnet. Mit der Hilfe einer analytischen Funktion werden anschließend die Anteile aller Mitwirkenden addiert. Dies wird analog für alle Bearbeitungen im Datensatz durchgeführt ist. Die Ergebnisse werden in einem Diagramm visualisiert, in dem die Anzahl der Mitwirkenden den kumulierten Prozentangaben gegenübergestellt wird. Zusätzlich wird die Anzahl der Mitwirkenden berechnet, die für 98 % aller erstellten Objekte und Objekt-Bearbeitungen verantwortlich sind. Diese Schwelle wird in Anlehnung an die eingangs erwähnte „*Participation Inequality*“ definiert, da in Online-Communities wie OSM, der größte Teil der Mitwirkenden insgesamt nur für einen sehr geringen Teil der Daten verantwortlich ist. Eine vollständige Gleichverteilung der Beiträge über alle Mitwirkenden hinweg ist dann gegeben, wenn beide berechnete Kurven einer 45° Linie entsprechen. Im Umkehrschluss ist die Verteilung umso ungleichmäßiger, je weiter sich die Kurven von dieser entfernen. Ersteres zeugt von einer durchgehend gleich hohen Menge an erstellten Objekten jedes einzelnen Mitwirkenden. Je weniger Mitwirkende für den größten Teil der erstellten Daten verantwortlich sind, desto höher ist die Abhängigkeit von diesen wenigen in einem Gebiet in OSM. Gleiches gilt für die akkumulierte Zahl der Objekt-Bearbeitungen für den Fall, dass ein Programm automatisiert Daten bearbeitet. Datenimporte haben eine besonders große Auswirkung, da sie zumeist von einem Mitwirkenden vollzogen werden und somit eine signifikante Auswirkung auf diese Analyse haben. Aus diesem Verfahren kann demnach abgeleitet werden, dass eine gleichmäßige Streuung der Beiträge (erstellte Objekte und Objekt-Bearbeitungen) über die Community generell als besser einzustufen ist. Die 98 %-Schwelle verdeutlicht zudem, wie viele Mitwirkende für 98 % der erstellten Daten bzw. Daten-Bearbeitungen verantwortlich sind.

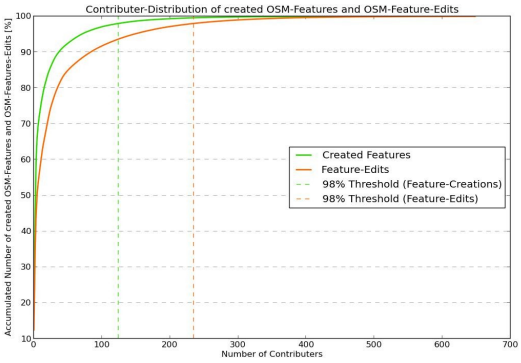

Abb. 3: Anzahl an Mitwirkenden & der kumulierte Prozentsatz an erstellten OSM-Objekten (grün) & OSM-Objekt-Bearbeitungen (orange) in Heidelberg

Die Ergebnisse der beispielhaften Analyse für das Gebiet von Heidelberg werden in Abbildung 3 in zwei Kurven visualisiert. Es wird dabei deutlich, dass bei den erstellten OSM Objekten eine größere Ungleichheit vorzufinden ist als bei der Anzahl an Objekt-Bearbeitungen. Der anfänglich steile Anstieg beider Kurven, der bei den erstellten Objekten stärker ausfällt, deutet auf einige wenige Mitwirkende hin, die für einen deutlich höheren Anteil verantwortlich sind.

3.4 Beitragsverhalten der Mitwirkenden

Wie bereits erwähnt, spielt die Aktivität der Mitwirkenden eine große Rolle. Folglich ist es wichtig, die Mitwirkenden zusätzlich anhand von ausgewählten quantitativen Parametern zu charakterisieren. Hierzu sind mehrere wissenschaftliche Studien zu finden, die in unterschiedlicher Art und Weise darauf eingehen (MOONEY & CORCORAN 2012, MOONEY & CORCORAN 2012a, NAPOLITANO & MOONEY 2012, NEIS & ZIPF 2012, REHRL ET AL. 2012). Durch die unterschiedlichen Beitragsintensitäten und -arten können allgemeine statistische Aussagen zum Beitragsverhalten von einzelnen oder von Gruppen von Mitwirkenden in einem Gebiet getroffen werden. Die folgenden Punkte definieren dabei die einzelnen Parameter für jedes aktive Mitglied:

- Anzahl der erstellten Punkte, Linien und Polygone,
- Anzahl der Bearbeitungen an Punkten, Linien und Polygonen,
- Anzahl der gelöschten Punkte, Linien und Polygone,
- Anzahl der Punkte, Linien und Polygone, an denen der Mitwirkende Attribute bearbeitet hat,
- Anzahl der Bearbeitungen, bei denen der Mitwirkende Attribute gelöscht hat ohne neue hinzuzufügen,
- Anzahl der Bearbeitungen, bei denen der Mitwirkende Attribute hinzugefügt hat ohne welche zu löschen,
- Anzahl der unterschiedlichen jemals verwendeten OSM Attribute,
- Anzahl der verwendeten Attribute, die lokales Wissen bzw. ein "Vor-Ort-sein" des Mitwirkenden erfordert,
- Anzahl der insgesamt bearbeiteten Geometrien (Hinzufügen, Löschen oder Verschieben von Punkten),
- Anzahl der erstellten invaliden Geometrien,
- Anzahl der unterschiedlichen jemals bearbeiteten Punkten, Linien und Polygonen.

Diese Eigenschaften werden quantitativ berechnet. Bei Aussagen zur Bearbeitung von Attributen durch die Mitwirkenden werden die Ergebnisse dadurch ermittelt, dass konsekutive Versionen aller Punkte, Linien oder Polygone miteinander verglichen werden. Aus der Differenz zueinander können anschließend Aussagen zu Veränderung der Attribute getroffen werden. Selbiges Verfahren wird auch zur Ermittlung von hinzugefügten oder entfernten Punkten in Geometrien angewendet. Für die Bestimmung von Objekten mit Attributen, die lokales Wissen bzw. ein "Vor-Ort-sein" der Mitwirkenden erfordern, wird auf den Ansatz von NAPOLITANO & MOONEY (2012) zurückgegriffen. Dieser wird jedoch dahingehend abgeändert, dass andere als die von den Autoren vorgeschlagenen OSM Attribute verwendet werden, da diese zu restriktiv und zu selten verwendet erscheinen.

4 Ausblick

Die in diesem Artikel vorgestellten Verfahren zeigen, auf welche Art und Weise in einem intrinsischen Ansatz Informationen über Mitwirkende des OSM Projektes generiert und Bewertungen auf Grundlage der Anzahl der Mitwirkenden, deren Aktivität und deren Anzahl an Beiträgen getroffen werden können. In diesem Kontext treten jedoch einige Fragen auf, die Gegenstand weiterer Forschung sein könnten. Wie angedeutet wurde, sind aus der Datenhistorie umfangreiche Profile generierbar, die auf der Basis der Beiträge der Mitglieder berechnet werden können. Diese Informationen könnten mit OSM Objekten verknüpft werden, um neuartige Aussagen zur Qualität der Daten zu erhalten. Hierfür müssten Profile erstellt werden, die Aussagen über die Vertrautheit der Mitwirkenden mit dem Analysegebiet oder dem OSM Projekt insgesamt enthalten. Anhand dieser könnten in der Folge Rückschlüsse auf gute bzw. schlechte Datenqualität an Objektkategorien oder Gebieten gezogen werden.

Danksagung

Die Autoren bedanken sich bei allen, die direkt oder indirekt zu diesem Artikel beigetragen haben, insbesondere der GIScience Research Group der Universität Heidelberg und bei allen Mitwirkenden des OSM-Projektes. Diese Arbeit wurde teilweise durch die Klaus-Tschira Stiftung (KTS) Heidelberg finanziert.

Literatur

CIEPLUCH, B., MOONEY, P. & WINSTANLEY, A. C. (2011), Building Generic Quality Indicators for OpenStreetMap. In: Proceedings of the 19th annual GIS Research UK GISRUK. Portsmouth, England.

GIRRES, J.-F. & TOUYA, G. (2010), Quality Assessment of the French OpenStreetMap Dataset. In: Transactions in GIS, 14 (4), S. 435 - 459.

HAKLAY, M. (2008), How good is Volunteered Geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. In: Environment and Planning B: Planning and Design, 37 (4), S. 682 - 703.

OPENGEODATA.ORG (2013), 1 million OpenStreetMappers. http://opengeodata.org/1-million-openstreetmappers (01.02.2013).

REHRL, K., GRÖCHING, S., HOCHMAIR, H., LEITINGER, S., STEINMANN, R. & WAGNER, A. (2012), A conceptual model for analyzing contribution patterns in the context of VGI. In: LBS 2012 - 9th Symposium on Location Based Services. Berlin: Springer.

LIEBERMANN, M. & LIN, J. (2009), You Are Where You Edit: Locating Wikipedia Contributors Through Edit Histories. In: Proceedings of the Third International ICWSM Conference (2009).

MOONEY, P., CORCORAN, P. & WINSTANLEY, A. C. (2010), A study of data representation of natural features in openstreetmap. In: Proceedings of the 6th GIScience International Conference on Geographic Information Science, GIScience 2010, University of Zürich.

MOONEY, P. & CORCORAN, P. (2011), Accessing the history of objects in OpenStreetMap. Proceedings of the 14th AGILE International Conference on Geographic Information Science, Utrecht 2011.

MOONEY, P. & CORCORAN, P. (2011a), The Annotation Process in OpenStreetMap. In: Transactions in GIS 16 (4), S. 561 - 579.

MOONEY, P. & CORCORAN, P. (2012), Characteristics of Heavily Edited Objects in OpenStreetMap. In: Future Internet, 4, S. 285 - 305.

MOONEY, P. & CORCORAN, P. (2012a), How social is OpenStreetMap? The 15th AGILE International Conference on Geographic Information Science. Avignon. April 2012.

MOONEY, P. & CORCORAN, P. (2012b), The Role of Communities in Volunteered Geographic Information Projects. Accepted peer-reviewed full paper: Proceedings of the 9th Sym posium on Location Based Services, Munich, Germany - October 2012. Springer's Lecture Notes in Geoinformation and Cartography.

NAPOLITANO, M. & MOONEY, P. (2012), MVP OSM: A Tool to identify Areas of High Quality Contributor Activity in OpenStreetMap. The Bulletin of the Society of Cartographers. Summer 2012.

NIELSON, J. (2006), Participation Inequality: Encouraging More Users to Contribute. http://www.nngroup.com/articles/participation-inequality/ (11.01.2013).

NEIS, P., ZIELSTRA, D. & ZIPF, A. (2012), The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007-2011. In: Future Internet, 2012 (4), 1-21.

NEIS, P., M. GÖTZ & A. ZIPF (2012a), Towards Automatic Vandalism Detection in OpenStreetMap. ISPRS International Journal of Geo-Information. 2012, 1(3), 315-332.

NEIS, P. & ZIPF, A. (2012), Analyzing the Contributer Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap. In: ISPRS International Journal of Geo-Information, 1, S. 146 - 165.

VAN EXEL, M., DIAS, E. & FRUIJTIER, S. (2010), The impact of crowdsourcing on spatial data quality indicators. In GIScience 2010: Proceedings of the 6th International Conference in Geographical Information Science. University of Zurich.

ZIELSTRA, D. & ZIPF, A. (2010): A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. AGILE 2010. The 13th AGILE International Conference on Geographic Information Science. Guimarães, Portugal.