Exposé zur Bachelorarbeit

Konstantin Bruckert - MatrNr. 558290 - WiSe 2019/20

Titel 1: Potenzial öffentlicher Feinstaub-Sensordaten in Berlin am Beispiel OpenSenseMap

Titel 2: Potenzial und Limitierung offener Sensordaten am Beispiel Berliner Feinstaubmessungen der OpenSenseMap

Inhalt

Einleitung

Fridays for Future, Flugscham, Dieselverbot und Abgas Skandal – Umweltschutz und eine damit einhergehende Luftverschmutzung bestimmt heutzutage den Diskurs öffentlicher Debatten. WHO, EU und Umweltbundesamt ringen um Belastungsgrenzwerte und sind sich zugleich einig, dass Menschen insbesondere in Städten, vor zu intensiver Exposition geschützt werden müssen. Aufbauend auf den Arbeiten und Untersuchungen des Projektstudiums „OpenSensorData“ soll eine Bachelorarbeit entstehen, in der Feinstaubwerte von Berliner Sensoren offener Quellen am Beispiel „OpenSenseMap“ untersucht werden. Ziel der Untersuchung ist es eine möglichst genaue Aussage zu treffen, inwieweit und auf welchem Niveau diese Feinstaubsensoren zur Verbesserung der allgemeinen Informationslage beitragen können.

Auf der Plattform OpenSenseMap können Nutzer mit Zugang zu einem Sensor dessen Daten veröffentlichen und diese anderen Besuchern offen und frei zur Verfügung stellen – auf Wunsch auch anonym. Dank dieser „Citizen Scientists“ entsteht eine beachtliche Menge unterschiedlichster Messwerte mit entsprechenden Zeit- und Ortsangaben, die mittels einer offenen und leicht zugänglichen API auch maschinell erfasst werden können.

Teil 1: Aufbau einer Datensammlung von Sensordaten

Ziel: Softwaremodul “Datensammlung”, macht Feinstaubwerte von der OpenSenseMap und etwaiger Drittanbieter dahingehend ansprechbar, sodass spätere Teilabschnitte der Arbeit damit arbeiten können.

Im ersten großen Teilschritt der Bachelorarbeit soll versucht werden, die Daten von der Plattform zu extrahieren und ansprechbar zu machen. Hierbei soll idealerweise ein Softwaremodul entstehen, dass die Anfragen an die OpenSenseMap-API so gut bündeln kann, dass sich die Notwendigkeit einer eigenen Datenspeicherung erübrigt. Sollte dies aufgrund technischer oder formaler Limitierungen seitens der API (Bsp: Langsame Server, Zugriffslimit) nicht möglich sein, muss ein alternatives Konzept mit eigener

Datenverwaltung oder begrenzten Datenumfang erstellt werden. Auch alternative Datenquellen, wie z. B. die staatlicher Stellen oder privater Unternehmen, die im späteren Verlauf der Arbeit benötigt werden (z. B. bei der Datenvalidierung), sollen in diesem Schritt integriert und ansprechbar gemacht werden. Hier werden auch anders geartete Sensordaten berücksichtigt, z. B. die Messwerte des BLUME Netzes, des Umweltbundesamtes oder Berechnungen privater Organisationen wie z. B. Breezometer.

Datenarten: Allgemein gibt es viele Messwerte, die auf den eingangs genannten Quellen zur Verfügung gestellt werden. Bei Schadstoffen, die unsere Gesundheit beeinflussen handelt es sich bis auf Feinstaub um Gase (Ozon und verschiedene *-Dioxide). Feinstaub hingegen ist ein Fest- oder Flüssigkörperschadstoff, der in großem Ausmaß auf menschliches Handeln zurückzuführen ist und deswegen für diese Arbeit als besonderer Fokus ausgegeben wird. Erkenntnisse sollen aber auf mögliche Generalisierungsansätze hin untersucht werden und es soll ein kleiner Ausblick entstehen, ob und inwieweit diese Erkenntnisse auf andere Luftschadstoffe übertragen werden könnten.

Bisherige Arbeiten im Modul “OpenSensorData” hatten den besonderen Fokus auf Daten von in Berlin aufgestellten Feinstaubsensoren. Zunächst erscheint dieser Ansatz legitim, da er ein so globales Phänomen wie Luftverschmutzung im Alltag greifbar macht und die potenziellen Interessenten besonders mit Informationen lokaler Natur angesprochen werden können. Dieser Ansatz – den Datenraum der Messstationen auf Berlin zu beschränken – soll grundsätzlich beibehalten werden. Bei statistischen Verfahren führen weniger Daten jedoch meist zu unpräzisen Modellen. Deshalb soll im Laufe der Arbeit eine Untersuchung gemacht werden, welche Definition der Grunddatenmenge (Gebiet, Zeitraum, Intervall, Messwerte) für das spätere Modell besonders geeignet sind um es auf Aussagen für Berliner Sensordaten zu trainieren.

Teil 2: Entwickeln von Möglichkeiten zur Validierung von Sensordaten

Ziel: Softwaremodul “Modellrechnung”, dass mittels statistischer Verfahren Daten aus der OpenSenseMap aufbereitet, kalibriert, Fehler eliminiert und mögliche Zusatzinformationen zur Verbesserung mit einbezieht.

Im folgenden Teil der Arbeit sollen die nun aufbereiteten Daten mit einem geeigneten mathematischen bzw. statistischen Modell untersucht und ausgewertet werden. Die Implementierung soll in einem weiteren Softwaremodul stattfinden, dass abgestimmt mit dem aus Teil 1 zusammenarbeitet. Ein beispielhafter Aufbau dieses Modells könnte wie folgt aussehen und erweitert werden:

- Ausreißer eliminieren: Grobe Unstimmigkeiten in den Daten werden durch Vergleiche innerhalb des Datensatzes und durch Plausibilitätsprüfungen eliminiert.
- Daten glätten: Sensordaten werden kalibriert und normalisiert. Es wird versucht Ansätze zu finden, Fehlermuster in den Messwerten zu erkennen und diese bestmöglich aus den Daten herauszurechnen.
- Kreuzvergleiche mit kalibrierten Daten: Wie verhalten sich die Messwerte im Vergleich zu anderen Messtechniken (private/staatliche Messungen)? Können die

Daten mithilfe anderer Messtechniken robuster gemacht werden, bzw. Fehler behoben?

- Gewinnen von Zusatzinformationen: Gibt es Verfahren mit denen die Daten erweitert werden können? Durch könnten z.B. Messwerte für unbekannte Orte entstehen oder eine Heatmap als Indikator. Könnte mit anderen Daten, die im kausalen Zusammenhang mit Feinstaubmessungen stehen, zusätzliche Möglichkeiten der Validierung oder Vorhersage entstehen (Fiktive Beispiele: starker Wind =? weniger Feinstaub oder hohe Benzinpreise =? mehr Feinstaub). In diesem Teilabschnitt soll ein Versuchslabor entstehen, das eine interaktive Komponente miteinbringt und die Kreativität des Studenten im Umgang mit Daten widerspiegeln soll – vielleicht sogar mit neuen Forschungsergebnissen.

Wie die Software-Implementierung des Modells konkret aussieht, welche Techniken bzw. Rechenmodelle genutzt werden können, um den größtmöglichen Informationsgewinn zu erlangen, soll in der Arbeit evolutionär entwickelt werden.

Teil 3: Aufbereitung und Evaluation der Daten

Ziel: Statistische Auswertung der Daten mit schriftlicher Abhandlung der Erkenntnisse aus der Arbeit und abschließender Bewertung.

Aufgrund der extrem großen Datenmenge kommt es schnell zu unüberschaubaren Strukturen. Zwar können Erkenntnisse z. B. aus der Betrachtung einzelner Messstationen gezogen werden, globale Phänomene oder Muster lassen sich aber nur schwer auf den ersten Blick erkennen und müssen ermittelt werden. Zunächst soll in diesem Teilabschnitt eine Aufbereitung der Daten stattfinden. Diese werden auf statistische Muster untersucht, um eine zentrale Frage der Arbeit zu beantworten: Unter der Annahme, dass staatliche oder private Messwerte korrekt sind, besitzen die analysierten Daten der OpenSenseMap eine statistische Signifikanz? Von dieser Frage ausgehend, sollen die Daten und Erkenntnisse weiter ausgewertet werden um statistische Aussagen über die Sensoren, Messwerte und die OpenSenseMap zu formen.

Hierbei werden die Daten zunächst vor und nach der Validierung/statistischen Bereinigung verglichen und auf ihre Präzision bzw. Unstimmigkeit untersucht. In einem weiteren Schritt soll insbesondere auf die Limitierungen der Messwerte eingegangen werden: Gibt es eine Erwartungshaltung, die die Sensoren auf keinen Fall erfüllen? Verkäufer, Plattformbetreiber und User/Bürger haben Erwartungshaltungen an die Messwerte – gibt es hier große Differenzen zwischen Erwartung und Realität? Abschließend soll untersucht werden, welches Potenzial in den Messwerten steckt: Welche Aussagen können die Daten mit welcher Wahrscheinlichkeit wie Präzise treffen? Welche Möglichkeiten bestehen, die Daten statistisch oder durch technische Verbesserungen am Messverfahren aufzuwerten?