Entwicklung und Analyse eines Modells zur raum-zeitlichen Modellierung von Umweltinformationen auf Basis von Crowdsourcing

Pierre Karrasch, Daniel Henzen und Matthias Müller
Technische Universität Dresden · pierre.karrasch@tu-dresden.de

Short paper

1 Einleitung

Die flächenhafte Modellierung von Umweltinformationen, wie meteorologischen Daten und Luftschadstoffinformationen, basiert im Wesentlichen auf den Daten administrativer Messnetze, die eine nur vergleichsweise grobe räumliche und zeitliche Auflösung aufzuweisen haben. Dazu gehören beispielsweise Messnetze der Landesämter sowie des Deutschen Wetterdienstes (DWD). Insbesondere in urbanen Umgebungen sind daraus abgeleitete flächenhafte Modellierungen vielfach nur eingeschränkt geeignet. Die starke Heterogenität dieser Räume, die einen direkten Einfluss auf die Messungen oben genannter Parameter haben kann, kann mithilfe dieser Verfahren nur unzureichend abgebildet werden. Die persönliche Exposition der Bürger im urbanen Raum gegenüber diesen Parametern kann dadurch stark gegenüber mittleren Konzentrationen für den Großraum eines urbanen Gebietes abweichen.

An dieser Stelle setzt ein Modellentwurf an, der mithilfe von Citizen Science und Crowdsourcing-Methoden eine zusätzliche Datenquelle generiert und nutzt, um einerseits der Heterogenität städtischer Räume gerecht zu werden und andererseits auch genau dort Daten zu erheben, wo Bürger diesen Umweltparametern ausgesetzt sind. Dazu werden Low-Cost-Messstationen entwickelt und eingesetzt, deren Basis exemplarisch in Abb. 1 dargestellt ist.

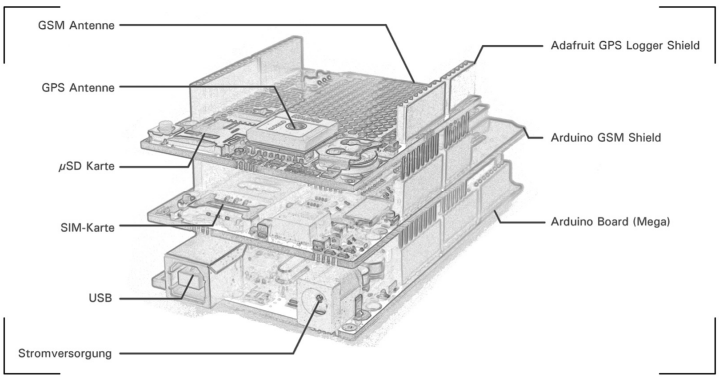

Abb. 1: Grundaufbau einer Low-Cost-Sensorstation basierend auf der Arduino Technologie

AGIT – Journal für Angewandte Geoinformatik, 1-2015, S. 409-414. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-557-7, ISSN 2364-9283, eISSN 2509-713X, doi:10.14627/537557056.

Die Messstationen bestehen im Wesentlichen aus einem Arduino Uno/Mega Board, einem GPS-Shield zur Bestimmung der Position der Messungen sowie einem GSM-Shield zur Speicherung und Übertragung und der Messwerte via Mobilfunknetz. Die gewählte Technologie ermöglicht den weitgehend flexiblen Einsatz unterschiedlicher Sensoren. Die im Folgenden vorgestellte Konzeption und die daraus abgeleiteten Ergebnisse nutzt als Anwendungsfall die Daten eines DHT22-Sensors, der Informationen über die Temperatur und Luftfeuchtigkeit zur Verfügung stellt.

2 Konzeption

Die grundlegende Idee des entwickelten Modells basiert auf der Annahme, dass jeder Ort ein für ihn spezifisches und charakteristisches Verhalten gegenüber einem speziellen Umweltparameter hat und sich dieses Verhalten abhängig vom betrachteten Zeitpunkt ändern kann. Damit kann der Messwert an einem konkreten Ort als Funktion seiner charakteristischen Umgebung, dem Zeitpunkt der Datenerhebung bzw. eines Zeitintervalls und einem verbleibenden zufälligen Term (Random) verstanden werden.

Kann weiterhin davon ausgegangen werden, dass ein administratives Messnetz vorhanden ist, für deren Stationen die formulierten Abhängigkeiten ebenfalls unterstellt werden können, sollte es möglich sein, einen funktionalen Zusammenhang zwischen Messungen an einem beliebigen Ort und einer oder mehrerer offizieller Messstationen herzustellen, der sich jedoch dynamisch im Laufe der Zeit ändern kann. Ist diese mathematisch beschreibbare Beziehung bekannt, besteht die Möglichkeit, für den betrachteten Ort und ohne kontinuierlich auf reale Messdaten zurückgreifen zu müssen, einen Messwert zu modellieren. Abb. 2 zeigt den schematischen Aufbau der Konzeption dieses Modellansatzes.

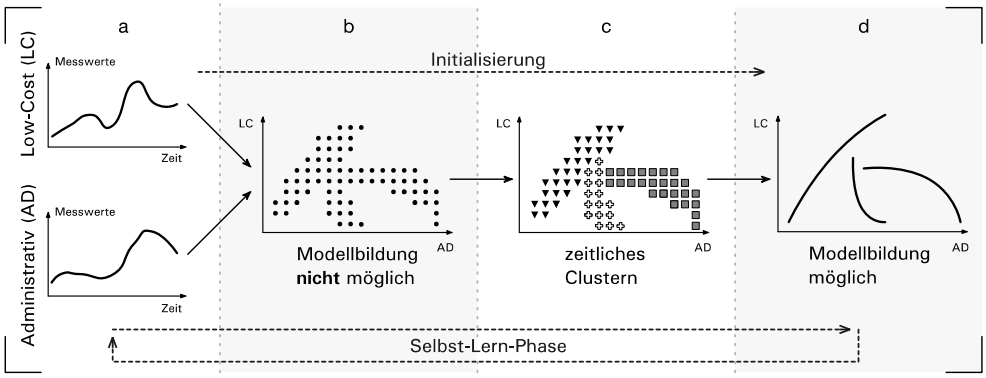

Abb. 2: Schematischer Workflow zur Bildung von statistischen Modellen zur charakteristischen Beschreibung eines Ortes auf Basis von Low-Cost-Sensor-Messungen und administrativen Messungen

Der Aufbau des Modells vollzieht sich im Wesentlichen über zwei Arbeitsschritte; einer Initialisierungsphase und einer daran anschließenden Selbst-Lern-Phase.

2.1 Initialisierungsphase

Beginnend mit der Initialisierungsphase erfolgt eine detaillierte Analyse des funktionalen Zusammenhangs auf der Grundlage erster Messungen in einem definierten räumlichen Abschnitt bzw. an einer statischen Position im Raum. Diese Daten können durch Crowdsourcing erhoben sein, oder wie im vorliegenden Fall, durch permanente Messungen einer lokalen Low-Cost-Station, ersetzt werden (vgl. Abb. 2a).

Datenpaare entstehen durch die Zuordnung von eigenen erhobenen Daten zu zugehörigen Messwerten einer administrativen Messstation, die die Grundlage für ein erstes mathematisches Modell bilden. Dabei erfolgt die Modellbildung, indem die Crowdsourcing Messungen als abhängige Variable und die administrativen Messungen als unabhängige Variable fungieren. Aufgrund der Tatsache, dass für beide Arten von Messungen der gleiche Parameter gewählt wurde und die administrative Messstation, zu der das Modell referenziert wird, einen räumlichen Zusammenhang aufweist, kann in einem ersten Schritt von einem linearen Ansatz ausgegangen werden, wenn nicht, wie in Abb. 2b dargestellt, weitere (unbekannte) Einflüsse den funktionalen Zusammenhang beeinflussen.

Die darauf folgenden detaillierten statistischen Analysen ermöglichen eine Beschreibung der Modellqualität in Form der Standardabweichung der Modellparameter, der Standardabweichung der Punktprognose sowie der Signifikanz. Gleichzeitig sind sie ein Indikator dafür, ob der gewählte Modellansatz eines linearen Modells geeignet ist. Sollte das nicht der Fall sein, kann dieser durch seine Modellstruktur in Form von beispielsweise polynomialen oder exponentielle Ansätze erweitert werden.

Zu diesem Zeitpunkt wird davon ausgegangen, dass die in Kapitel 2 erwähnte Charakteristik, allein durch einen funktionellen Zusammenhang hinreichend genau beschrieben werden kann und damit in seiner Struktur zeitinvariant ist. Abhängig von der Anzahl und zeitlichen Verteilung der verfügbaren Messungen können oder müssen geclusterte Modelle generiert werden, mit dem Ziel, die Qualität modellierter Messdaten weiter zu verbessern. Diese Clusterung kann zeitlich erfolgen, wie in Abb. 2c dargestellt, oder in bestimmte Messwertbereiche.

Entsprechend der gewählten Vorgehensweise, stehen am Ende der Initialisierungsphase mindestens ein oder eine Schar von mathematischen Modellen zur Verfügung, die die Charakteristik eines Ortes bezüglich einer administrativen Messstation initial beschreiben (vgl. Abb. 2d).

2.2 Selbst-Lern-Phase

Nach der Initialisierungsphase stehen die zeitlich oder/und nach Messwertbereichen geclusterten Modelle für die Modellierung der entsprechenden Umweltinformation ausschließlich auf Basis der Messwerte der Referenzstationen zur Verfügung und können in Abhängigkeit der gleichzeitig ermittelten Modellqualität genutzt werden, um räumlich und zeitlich Umweltmessdaten zu modellieren.

Die nun folgende Selbst-Lern-Phase nutzt wiederum die durch Crowdsourcing in unregelmäßigen Abständen neu verfügbaren realen Messungen. Damit wird es möglich, das in der Initialisierungsphase generierte Modell zu kontrollieren und gegebenenfalls die Modellparameter zu verbessern. Darüber hinaus ermöglicht diese Phase aber auch die Anpassung

der inhärenten Struktur (linear, polynomial, exponentiell, etc.) des Modells, falls dies die neuerlich erhobenen Daten notwendig machen. Gleiches gilt für Art und Umfang des durchgeführten Clusterings.

Unabhängig von der Verbesserung der Modellierung aktueller Umweltparameter ermöglicht der vorgestellte Ansatz auch die Durchführung von erneuten Reanalysen. Diesen können sich einerseits auf den bereits modellierten Zeitraum erstrecken, aber auch in den zeitlichen Bereich vor der Initialisierungsphase reichen. Dies ist möglich, so lange die Charakteristik des zu modellierenden Ortes und des Ortes der Referenzstation als konstant angenommen werden kann.

3 Technische Umsetzung

Wie bereits in Kapitel 1 beschrieben, können Low-Cost Messstationen je nach Anwendungsfall unterschiedlich konfiguriert werden. Es existieren verschiedene bereits vorgefertigte Module für unterschiedliche Szenarien wie der Speicherung von Daten auf SD (Backup), dem Aufnehmen von Positionsinformationen oder dem Versenden von Daten über Mobilfunk/GSM-Verbindungen. Für den oben beschriebenen Anwendungsfall besteht eine Messstation aus all den soeben aufgeführten Komponenten, zuzüglich Sensoren unterschiedlicher Qualitäten und Sensitivitäten für die Messung von Umweltparametern. Die Aufzeichnung und Veröffentlichung der Messdaten erfolgt in einem 5-Minuten-Takt und diese Daten werden nach der Übertragung an einen Dienst mittels GSM-Technologie durch einen Sensor Observation Service bereitgestellt. Die für die statistischen Analysen benötigten Messdaten der administrativen Stationen werden durch die entsprechenden Institutionen im Internet zur freien Verfügung gestellt. Durch entsprechend angepasste Dienste werden diese Daten bei Bedarf in das Analysesystem integriert. Allerdings können sie aufgrund fehlender Rechte nicht als Dienst zur Verfügung gestellt werden. Für die eigentlichen statistischen Analysen zur Modellbildung wird die freie Open-Source-Software R genutzt.

4 Ergebnisse

Die Umsetzung der in Kapitel 2 vorgestellten Modellkonzeption erfolgte auf der Basis der in Abbildung 1 dargestellten Sensorstation. Als Referenzstation diente eine Messstation des Deutschen Wetterdienstes (DWD, Station Dresden Strehlen) in einer Entfernung von ca. 3,5 km. Entsprechend der Konzeption erfolgte eingangs eine einfache lineare Modellierung des Zusammenhangs der Temperaturmessungen an beiden beteiligten Stationen (DWD, Low Cost). Das Ergebnis impliziert, dass ein linearer Ansatz grundsätzlich für die Modellierung dieses Zusammenhangs geeignet ist. Aber auch, dass die verbleibenden Residuen, die gleichzeitig ein Indikator für die Vorhersagegenauigkeit auf Basis der DWD sind, vergleichsweise hoch sind (vgl. Abb. 3, Standardabweichung: 2,1 K).

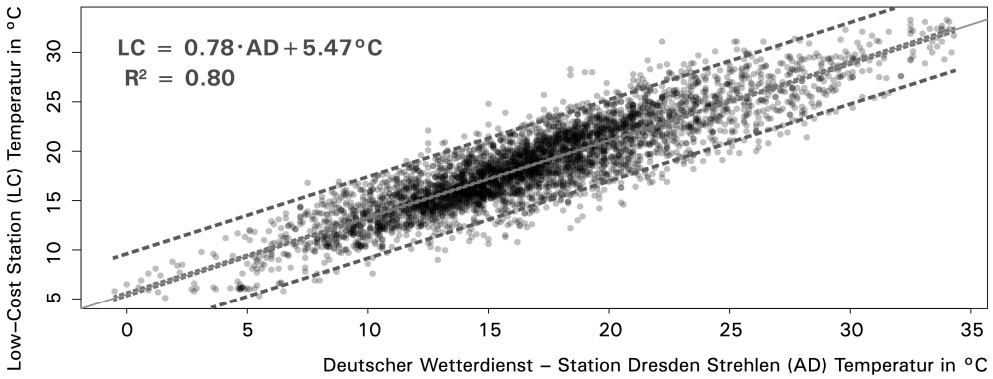

Abb. 3: Einfache lineare Regressionsanalyse zwischen administrativen Messwerten und Low-Cost-Sensor-Daten

Die statistischen Analysen der Modellierung weisen darauf hin, dass die Annahme eines einfachen linearen Zusammenhangs nicht aufrechterhalten werden kann. Zu diesem Zweck wurden die Messwertpaare in stündliche Intervalle geclustert.

Abbildung 4 zeigt beispielhaft diese Abhängigkeit des modellierten Parameters *Temperatur* in stündlichen Intervallen innerhalb eines Tageszyklus. Dabei zeigt sich im konkreten Fall, dass das Niveau (Achsenabschnitt) des linearen Modells zwischen 0 und 13 Uhr, zunächst additiv und danach subtraktiv, einer kontinuierlichen Schwingung unterliegt, während der Anstieg des Modells in diesem Zeitraum nahezu konstant ist. Dieses Verhalten ändert sich am Nachmittag ab 14 Uhr und hält bis etwa Mitternacht an. Die Modelle unterliegen nun einer Rotation mit einem Rotationspunkt bei etwa 10 °C. Der Mehrwert des Clusterings spiegelt sich insbesondere in der Verbesserung der Standardabweichung wider. Die Werte variieren nun zwischen 0,74 K und 1,4 K.

Die in dieser Phase der Modellbildung generierten statistischen Maße können ebenfalls als Kriterien für die Nutzbarkeit des Modellansatzes bezüglich eines gegebenen Anwendungsszenarios verwendet werden. Entspricht die Qualität der generierten Modelle den Anforderungen eines konkreten Szenarios, können nun allein auf Grundlage der Messwerte der DWD-Station, modellierte Werte für die korrespondierende Stelle im Raum zur Verfügung gestellt werden.

Die Analysen zeigten weiterhin, dass eine Anpassung der inhärenten Modellstruktur, beispielsweise durch eine polynomiale Modellierung, zu keiner weiteren Verbesserung der Modelle führte. Die Annahme eines grundsätzlich linearen, aber zeitlich dynamischen Ansatzes, konnte damit bestätigt werden.

Mit wachsender Datenverfügbarkeit lassen sich ebenfalls Aussagen über die räumliche Nutzbarkeit eines solchen *virtuellen Sensors* machen. Es ist davon auszugehen, dass je nach Grad der Heterogenität einer zu untersuchenden Umgebung der Wirkungsbereich der eruierten Modelle eingeschränkt ist.

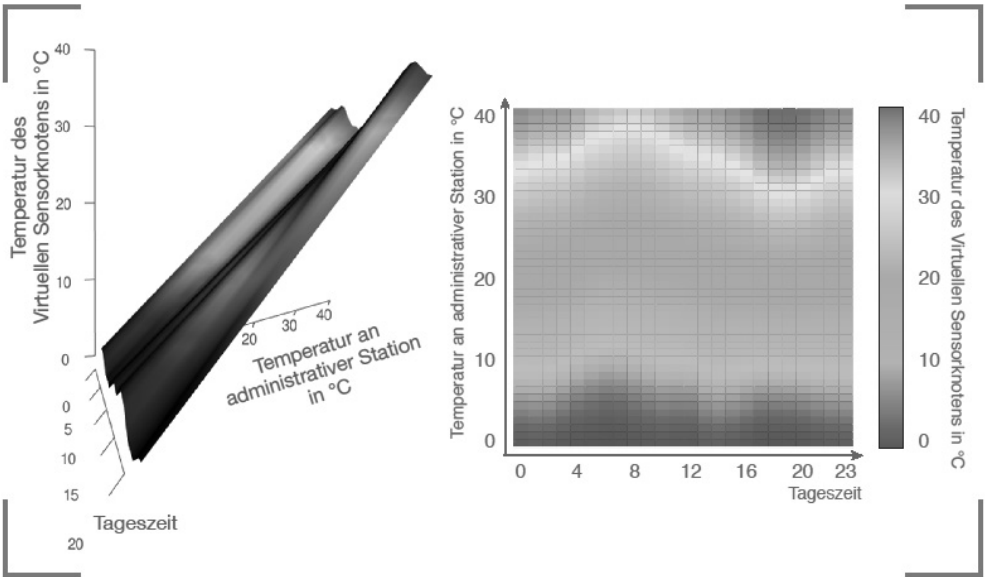

Abb. 4: Ergebnis der Initialisierungsphase für den Parameter Temperatur [°C] in zwei verschiedenen Perspektiven. Abhängigkeit des Messwertes (Crowdsourcing) von dem korrespondierenden Messwert einer administrativen Messstation (DWD) in stündlichen Zeitintervallen.

5 Zusammenfassung/Ausblick/Herausforderungen

Die dargestellte Methode zeigt, in welcher Form eine Modellierung und Verdichtung von Umweltparametern auf Grundlage bestehender Messnetze möglich sein kann. Letztendlich bleibt aber auch festzuhalten, dass die Verfügbarkeit von Daten das limitierende Element der Modellnutzbarkeit sein wird. Das bedeutet im Umkehrschluss, dass es nur dann einen erfolgreichen Einsatz geben kann, wenn die Etablierung des Modellkonzepts mit einer entsprechenden Motivation der potenziellen Crowd einhergeht.

Die praktische und flächenhafte Umsetzung wird Zeit benötigen; die resultierenden Ergebnisse anfangs nur als Patches verfügbar sein. Dennoch kann der Ansatz einen wertvollen Beitrag zur Modellierung von Umweltmessdaten, insbesondere in heterogen strukturierten Räumen leisten.

Danksagung

Die Finanzierung dieser Forschung erfolgte aus den Mitteln der Exzellenzinitiative des Bundes und der Länder (Zukunftskonzept der TU Dresden, Maßnahme „Support the best“).