



Flächennutzungsmonitoring VIII Flächensparen – Ökosystemleistungen – Handlungsstrategien

IÖR Schriften Band 69 · 2016

ISBN: 978-3-944101-69-9

Big Data und Data-Mining im Umfeld städtischer Nutzungskartierung

Bodo Bernsdorf, Julian Bruns

Bernsdorf, B.; Bruns, J. (2016): Big Data und Data-Mining im Umfeld städtischer Nutzungskartierung. In: Meinel, G.; Förtsch, D.; Schwarz, S.; Krüger, T. (Hrsg.): Flächennutzungsmonitoring VIII. Flächensparen – Ökosystemleistungen – Handlungsstrategien. Berlin: Rhombos, IÖR Schriften 69, S. 243-251.

Big Data und Data-Mining im Umfeld städtischer Nutzungskartierung

Bodo Bernsdorf, Julian Bruns

Zusammenfassung

Es ist festzustellen, dass die städtische Nutzungskartierung auf immer mehr Datenquellen zurückgreifen kann. Insbesondere handelt es sich um hochauflösende (Geo-)Daten von Fernerkundungsplattformen wie Satelliten aus dem Copernicus-Programm. Aber auch sogenannte Volunteer Geographic Information (VGI) spielen eine zunehmende Rolle. Speziell entwickelte Anwendungsprogramme, sogenannte „Apps“, kommen zum Sammeln solcher Rauminformationen in Frage. Und letztlich kommen Daten aus sozialen Netzwerken zum Tragen.

Dieser Beitrag beschäftigt sich mit der Anwendung von Big Data im geo-temporalen Umfeld: Daten mit großen Volumina, die immer schneller in den Prozess gelangen, aus unterschiedlichsten Quellen stammen, unterschiedliche Informationsgehalte aufweisen und mit Unsicherheit behaftet sind. Sie liegen möglicherweise nicht flächendeckend vor, bieten mannigfaltige Bodenaufösungen, sind lückenhaft – dies sind alles Aspekte, die den gängigen Kriterien für „gute“ Daten widersprechen. Man wünscht sich flächendeckende, hochauflösende und hochaktuelle Daten. Der Vorteil bei der Nutzung von Big Data liegt nicht in der „Güte“, sondern in der massenhaften Verfügbarkeit.

Der vorliegende Artikel ist als Werkstattbericht zu verstehen, der erste Ansätze in einem Anwendungsszenario zur Detektion sogenannter Intra-Urban Heat Islands, innerstädtischer Hitzeinseln, aufzeigt.

1 Einführung: Big Data und Data-Mining

1.1 Ausgangslage: Big Data

Immer größer werdende, heterogene Datenbestände lassen sich mit üblichen Methoden nicht mehr analysieren. Zum einen werden die Probleme zu groß um sie mit Rechenkapazitäten von klassischen Computern zu lösen. Zum anderen scheitern Verfahren an der neuen Komplexität der Daten, der sogenannte Fluch der Dimensionalität (Bellmann 1961). Weiterhin entstammen manche Daten potentiell unsicheren Quellen. Geographische Informationssysteme (GIS) können sie nicht fassen oder gar sinnvoll verarbeiten.

GIS wurden entwickelt um Toblers 1. Gesetz der Geographie abzubilden. Es beschreibt den Fakt, dass räumlich benachbarte Objekte sich hinsichtlich ihrer Attribute oft ähnli-

cher sind, als räumlich entfernte Objekte – das Prinzip der Autokorrelation (zitiert in Shekar, S. 2014). GIS können mit wenigen, sehr genau erhobenen Werten Interpolationen liefern und räumlich vorhersagen. Aber sie sind nicht in der Lage, komplexere Strukturen korrekt zu verarbeiten oder gar (Nah-)Echtzeit-Analysen auf Big Data durchzuführen. Als Big Data werden Daten interpretiert, wenn sie den sogenannten vier „V“ entsprechen. Das Datenvolumen (Volume) ist wesentlich, je nach Umgebung werden aber verschiedene Volumina als „big“ angesehen. Die Geschwindigkeit, mit der Daten in den Prozess gelangen (Velocity), erzeugt und transferiert werden ist ein weiteres Kriterium. Zudem entstammen Daten aus unterschiedlichsten Datenquellen und sind damit sehr verschieden etwa bezogen auf Erfassungsmethoden (Variety). Und schließlich weisen sie Unsicherheiten bezüglich ihres Inhalts, der Integrität und der Robustheit der Daten auf (Veracity).

Die üblichen Anforderungen an die Datenqualität – Aktualität, Flächendeckung und Auflösung – treffen auf Big Data nicht zu. Stattdessen liegen deutlich mehr und heterogenere Daten vor. Die jeweils enthaltene Information muss mit Hilfe von Data-Mining-Verfahren erst extrahiert und bewertet werden.

1.2 Mustersuche: Data-Mining

Komplexe, nicht-lineare Strukturen sind für ein klassisches GIS problematisch. Stattdessen müssen die Beziehungen lokal analysiert werden.

Diese Unterschiede zwischen lokalen und globalen Modellen beschreiben Shekhar, Zhang (2004). Betrachtet man eine Punktwolke global und legt eine Regression darüber, kann eine vollständig andere Aussage resultieren, als wenn man sich die lokalen Gruppen ansieht. Im Beispiel der Abbildung 1 resultiert im globalen Modell eine Regression mit positiver Steigung, in den beiden lokalen Modellen jeweils eine mit negativer Steigung.

Diese Erkenntnis ist abgeleitet aus dem 2. Gesetz der Geographie nach Goodchild (zitiert in Shekhar, Zhang 2004): Globale geographische Modelle können inkonsistent zu lokalen Modellen sein, vergleiche z. B. die Modelle Moran's I und LISA (Anselin 1995). Bei der Betrachtung geographischer Zusammenhänge, zum Beispiel einer Landnutzungskartierung via Satellitenbild-Klassifikation, kann man daher oft Trainingsgebiete nicht über zu große Räume anwenden.

Hier liegt die Motivation für den Einsatz von Data-Mining. Es geht darum, die Daten-Dimensionen soweit zu reduzieren, dass bekannte Algorithmen in einer angemessenen Zeit interpretierbare Ergebnisse liefern. Denn die heutigen Computer-Kapazitäten sind oft nicht in der Lage, Big Data-Bestände vollständig durchzurechnen. Im Rahmen einer Studie für den Deutschen Bundestag wurde dazu eine Definition erstellt (Bernsdorf et al. 2015):

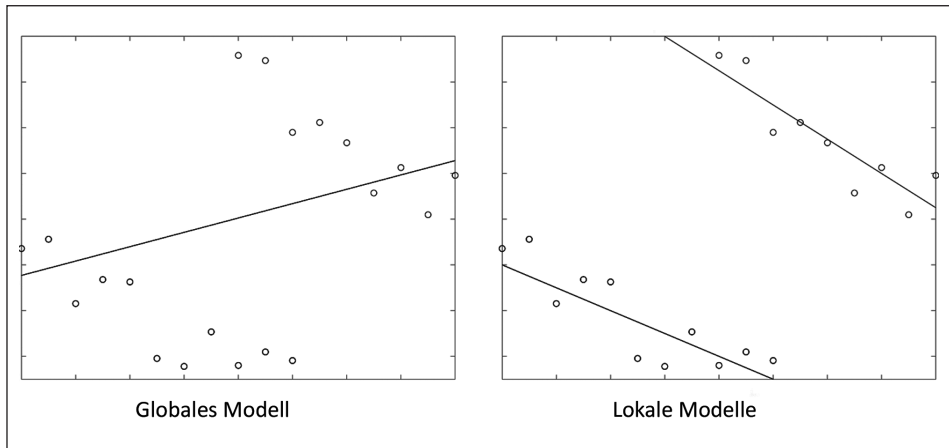


Abb. 1: Globales Modell vs. lokale Modelle (Quelle: Shekhar, Zhang 2004)

Unter Data-Mining mit Geodaten wird ein zielgerichteter Prozess verstanden, der durch Anwendung raumbezogener Methoden und mathematisch-statistischer Verfahren eine große, heterogene Geodatenmenge an der Grenze der technischen und algorithmischen Verarbeitbarkeit mit den aktuell verfügbaren Computer-Kapazitäten über auffällige Muster (Regeln oder Abhängigkeiten) derart reduzieren hilft, dass in angemessener Zeit die wesentlichen Erkenntnisse und Aussagen extrahierbar werden.

Dabei ist Data-Mining letztlich die Extraktion von Informationen aus Daten. Die statistischen Methoden sind durchaus bekannt. Mit der Ausreißer-Erkennung sucht man Daten, die stark von den restlichen Orten abweichen. In räumlichen Clustern gruppieren sich räumliche Muster nach Gesetzmäßigkeiten. Ko-Lokation erkennt, ob bestimmte Muster immer gemeinsam vorkommen. Sucht man auffällige Muster entlang eines Pfades in Raum und Zeit nutzt man die Sequenzanalyse. Die Assoziationsanalyse arbeitet klare Regeln für räumliche Muster heraus und in der Fernerkundung dient die Klassifikation dazu, Werte festen Klassen zuzuordnen, um etwa Landnutzungsmuster zu erkennen.

2 Das Forschungsprojekt BigGIS

Ziel des Projektes ist die Erstellung eines GIS zur Modellierung solcher komplexer, nicht-linearer Zusammenhänge und Entwicklungen in ständig wachsenden Mengen an unzuverlässigen, hoch-dimensionalen Daten, die Entwicklung neuer Mechanismen im Umfeld der Analytik und visuellen Analytik sowie die Nutzung von bestehendem Expertenwissen.

Evaluiert wird dieser Ansatz in drei Anwendungsszenarien, die alle einen Bezug zur Landnutzungsklassifikation haben.

2.1 Innerstädtische Wärmeinseln

Das vorgestellte Beispiel fokussiert die städtische Landnutzungsanalyse. Es behandelt die sogenannten „Intra-Urban Heat Islands“ (IUHI), also die innerstädtischen Hitzeinseln. Ein Schwerpunkt liegt auf der Identifikation der Verbreitung und Ausprägung städtischen Grüns als Grundlage für Verdunstungseffekte, die für Kühlung sorgen (Gill et al. 2007). Muster in der Grünausstattung sollen herausgearbeitet werden, um etwa (sozial-)räumliche Unterschiede belegen zu können. Eine Forschungsfrage ist, wie sich städtisches Grün in Verbreitung und Volumen erfassen lässt, um räumliche Differenzierungen abzuleiten. Diese Informationen lassen sich zur Identifikation und Vorhersage von innerstädtischen Hitzeinseln nutzen. Dies kann beispielsweise für die Streckenfindung mit der geringsten Wärmebelastung in der sommerlichen Stadt genutzt werden. Auch Frühwarnsysteme für Risikogruppen (Kindergärten, Altenheime etc.) können resultieren.

Hierzu wird ein Big Data Ansatz genutzt:

- ATKIS/ALKIS Daten (NoRA Baden-Württemberg),
- ein Thermalflug von Karlsruhe (Nachbarschaftsverband Karlsruhe),
- der Vegetationsindex NDVI (EnviSAT, Landsat),
- Sentinel 2,
- Kopter-erfasste Hyperspektraldaten,
- das LoD2-3D Daten der Stadt Karlsruhe,
- Werte von Messstationen des Deutschen Wetterdienstes (DWD) und des Landesamtes für Umwelt Baden-Württemberg (LUBW),
- Klimadaten vom Institut für Klimatologie und Meteorologie am Karlsruhe Institut für Technologie (IMK KIT), z. B. AERO-TRAM.

Diese Daten werden aktuell bereits genutzt. Geplant ist die Nutzung mobiler Messstationen, das sogenannte Participatory Sensing über Smartphone (Volunteer Geographic Information, gegebenenfalls auch Daten aus sozialen Netzwerken) oder Radar-Daten aus Befliegungen und/oder der Sentinel 1-Mission (C-Band-Radar).

Dabei entsprechen die Daten in großen Teilen den Kriterien für Big Data, wie sie oben vorgestellt wurden. Sie liegen rein volumenmäßig nicht im Petabyte-Bereich, aber es handelt sich durchaus um viele Terabyte an Daten (Volume). Wesentlich sind die anderen Kriterien. Mit der Sentinel-Mission könnten etwa alle fünf Tage neue Datensätze resultieren (Velocity), sie stammen aus unterschiedlichsten Quellen (Variety), die Bodenauflösung und die räumliche Abdeckung sind sehr verschieden und die Daten tragen große Unsicherheiten, da sie vollständig unterschiedlich erfasst werden (Veracity).

Die Vielzahl an Daten wird hierbei aus verschiedenen Gründen benötigt. Beispielsweise gibt es im Betrachtungsgebiet Karlsruhe lediglich eine offizielle Klima-Messstation des DWD, mit deren Daten langjährige Zeitreihen abzubilden sind. Zusätzlich existieren nur zwei weitere zuverlässige Messstationen. Jede dieser Stationen liefert zuverlässige Messungen an einem konkreten Punkt. Zur Vorhersage für eine Großstadt sind diese Daten nicht ausreichend. Demgegenüber stehen die Satellitendaten für eine rasterförmige Bodenauflösung und decken größere Bereiche ab. Ein potentieller Temperatur-Messwert, abgeleitet aus thermalen Informationen, steht integral für eine größere Fläche, die als Mischpixel gewertet werden muss und diverse Landbedeckungs- und -nutzungskategorien enthält (Abb. 2). Diese müssen nun anhand der exakten Messwerte der DWD-Stationen kalibriert und extrapoliert werden. Hieraus leitet sich eine technische und analytische Herausforderung ab, wie diese unterschiedlichen Quellen kombiniert werden können, um ein akkurates Bild für jeden Bereich der Stadt zu ermöglichen. Die durch die Datenlage entstehende Unsicherheit bei der Interpolation stellt eine zusätzliche Herausforderung dar.

Daher sieht die Vorgehensweise vor, dass eine Kombination aus Luft- und Bodenwerten von Messstationen genutzt wird und über zusätzliche Parameter extrapoliert werden

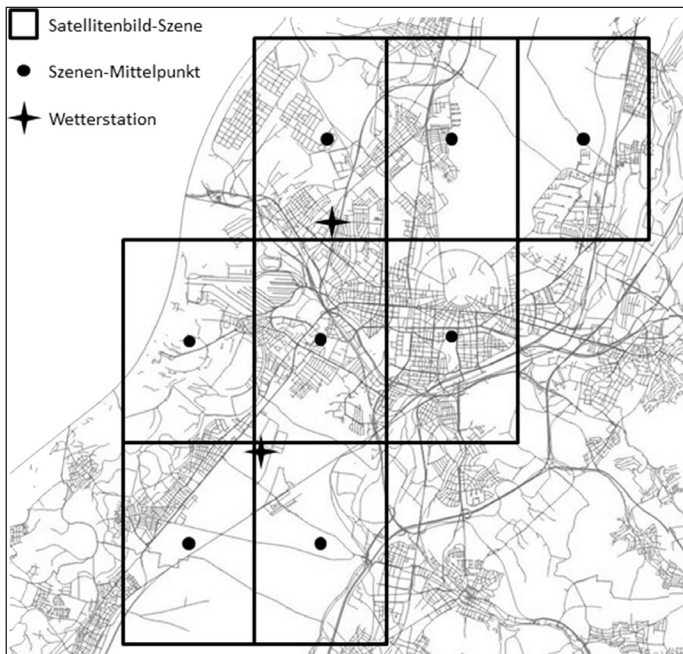


Abb. 2: Zwei Wetterstationen mit exakten, punktförmigen Temperatur-Messwerten im Vergleich mit der Abdeckung von EnvisAT-Thermalbilddaten. Letztere geben einen aus Thermalinformationen abgeleiteten „Temperaturwert“ an, der als Punktinformation für das Pixelzentrum gespeichert ist, aber integral die gesamte räumliche Ausdehnung des Mischpixels repräsentiert (Quelle: eigene Darstellung)

kann. Ein Beispiel ist der sogenannte Normalized Differenced Vegetation Index (kurz: NDVI) sowie die Bebauung. NDVI und Bebauung geben zusammen ein Bild über vermutlich wärmere oder kühlere Bereiche der Stadt. John Arnfield (2003) hat hierzu eine Zusammenfassung verschiedener Einflussfaktoren erstellt. Diese Information geht dann in die Extrapolation ein, um die Prognose sicherer zu gestalten. In einem ersten Schritt werden die Parameter zum gleichen Zeitpunkt vorhergesagt, um die Datenlage für zukünftige Prognosemodelle zu erweitern. Die dadurch entstehende Unsicherheit wird in späteren Modellen beachtet und dient als Grundlage für die Bestimmung neuer Messstandorte.

2.2 Methoden und Dateninput

Es wurden verschiedene, heterogene Datenquellen genutzt und anhand verschiedener Verfahren kombiniert. In einem ersten Ansatz wurden die Methoden der linearen Regression, Restricted Linear Regression sowie Bayes Hierarchical Modelling (BHM) genutzt, um die Temperatur an verschiedenen Orten vorherzusagen. Obwohl bislang nur vier Satelliten und die Distanz zum Messpunkt als Parameter in die Vorhersage einfließen, ergaben sich doch hohe Korrelationen von über 0,92 in allen drei Modellen. Jedoch enthielten die Vorhersagen – je nach Vorhersagepunkt – große Unsicherheiten. Insbesondere das BHM stellte diese Zusammenhänge besonders gut dar und ermöglichte eine genauere Analyse der Sachverhalte.

Im Projekt BigGIS werden weitere Datenquellen in Betracht gezogen. Um die Möglichkeiten von Participatory Sensing zu nutzen, werden insbesondere soziale Medien in Betracht gezogen. Verschiedene Verfahren für eine Auswertung wurden betrachtet, jedoch ist hier die große Herausforderung, die Daten räumlich zuzuordnen. Beim Kurznachrichtendienst Twitter weisen nur wenige Prozent der Nachrichten eine Geo-Referenz auf. Jedoch ist davon auszugehen, dass in solchen Nachrichten Hinweise bezüglich der sozialräumliche Differenzierung des städtischen Grüns und der Hitzebelastung vorhanden sind.

Eine bereits eingesetzte Datenquelle sind Daten aus Satellitenmessungen. Das EFTAS-Projekt DLM-Update sorgt für die Aktualisierung von Geobasisdaten. Es basiert auf der Idee, verfügbare ATKIS-Daten zu nutzen, um automatisch Trainingsgebiete abzuleiten. Dabei kann man davon ausgehen, dass über 95 % des Datenbestandes noch der Realität entspricht. Mit diesen Trainingsgebieten werden Landbedeckungsklassifikationen in aktuellen Satellitenbildern, etwa Sentinel 2, durchgeführt und eine Differenz zu den Objekten im Datenbestand errechnet. Hierdurch lässt sich eine Änderungsdetektion durchführen, die z. B. herausarbeiten kann, ob städtisches Grün nach wie vor im bekannten Umfang existiert (Abb. 3).

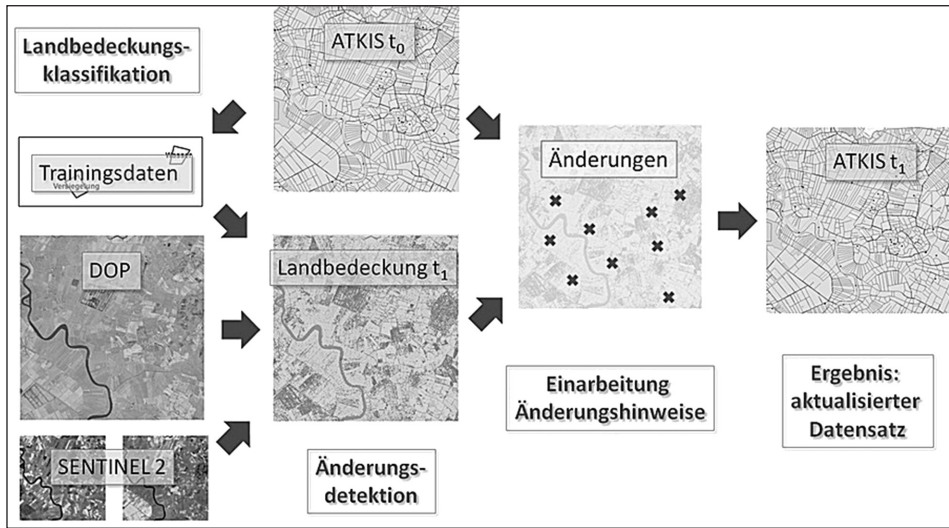


Abb. 3: DLM-Update-Prozess zur Veränderungsanalyse in städtischen Nutzungskartierungen (Quelle: eigene Darstellung des Prozessablaufs)

Letztlich soll das Angebot an Radardaten aus dem Copernicus-Programm eingesetzt werden. Ziel ist es, einerseits die räumliche Verbreitung des städtischen Grüns zu erfassen. Dieser Schritt erfolgt vorwiegend mit Sentinel 2-Daten. Andererseits muss das Volumen der Grünkörper abgeleitet werden. Denn nur hieraus können Verdunstungskapazitäten berechnet werden, die über den Kühlungseffekt wiederum Einfluss auf die Prognose der Hitzeinseln haben (vergleiche Gill et al. 2007). Geeignet ist dazu das X-Band, das von Satelliten wie TerraSAR-X und der Mission TanDEM-X bereitgestellt werden. Insbesondere letztere gibt Aufschluss über ein Digitales Oberflächenmodell.

Sentinel 1 ist als C-Band-Radar ausgelegt. Er gibt Informationen über die Verdunstungskapazität, da er den Vegetationskörper über die Feuchtigkeit unmittelbar erfasst. Und schließlich wäre ein L-Band-Radar für offene Flächen spannend, das zur Vorhersage

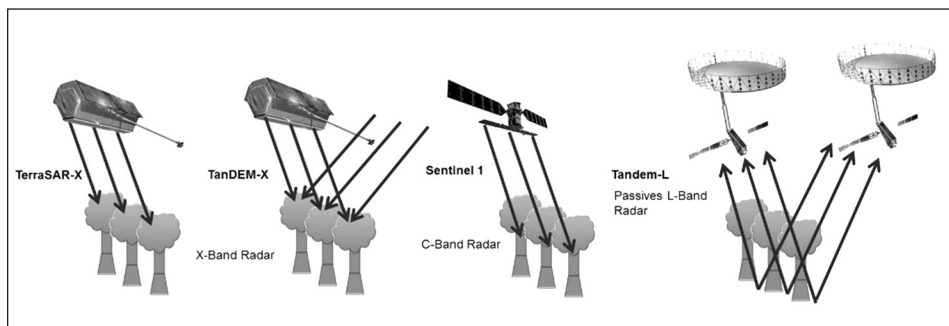


Abb. 4: Radar-Satellitenmissionen aus dem Copernicus-Umfeld (Tandem-L ist lediglich eine Konzeptstudie, Grafik eigener Entwurf; Quelle der Satellitenmodelle: DLR)

von Bodenfeuchte genutzt werden kann. Dringt L-Band durch einen Vegetationskörper, erhält man ein integrales Maß für die Feuchtigkeit in der Vegetation, ähnlich wie beim C-Band-Radar. Ein Nachteil ist, dass es aktuell keine kostengünstigen, nutzbaren L-Band-Satelliten gibt.

3 Fazit

Im Vergleich zur klassischen Erfassung von Nutzungsarten bietet Big Data vollständig neue Herausforderungen und Erkenntnisse. Es bedarf jedoch neuer Methoden, diese geo-temporalen Massendaten auszuwerten. Im Projekt BigGIS wird hierfür ein neuer Ansatz für Geographische Informationssysteme entwickelt. Das Beispiel der städtischen Nutzungskartierung zeigt Möglichkeiten auf, auch wenn es sich beim vorliegenden Artikel noch um einen Werkstattbericht – „work in progress“ – handelt. Es kann am Beispiel des Copernicus-Programms aufgezeigt werden, wie vielfältig Satellitendaten heute sind. Viele Fragestellungen werden damit fassbar. Die Städtische Grünausstattung wird als Anwendungsfall dargestellt, weil sie bezogen auf die IUHI dazu beitragen, Prognosen für die Temperatur und das Aufkommen solcher Hitzeinseln zuverlässiger als bisher zu gestalten. Auch der Ansatz Hyperspektraldaten, etwa von einer Satellitenplattform, zu erfassen, unterstützt solche Prognosen, denn Grünausstattung ist damit im Bereich zwischen 450 nm und 950 nm hervorragend über das Chlorophyll zu erfassen. Hyperspektrale Daten in Kombination mit Radardaten aus dem L-Band können in der Kombination auch eine witterungsbedingte Limitierung von Kühleffekten herausarbeiten, weil sich das Chlorophyll nach längeren Hitzeperioden bedingt durch die Trockenheit abbaut und Bodenfeuchtwerte zurückgehen. Weitere potentielle Datenquellen werden untersucht und evaluiert. Wie am Beispiel von Twitter jedoch ersichtlich ist, stellt hier insbesondere die Georeferenzierung Probleme dar.

4 Literatur

- Anselin, L. (1995): "Local indicators of spatial association – LISA." *Geographical analysis* 27.2, 93-115.
- Arnfield, A. J. (2003): Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. *Int. J. Climatol.*, 23, 1-26. doi:10.1002/joc.859
- Bellman, R. E. (1961): *Adaptive Control Processes*. Princeton University Press.
- Bernsdorf, B.; Bierbrauer, H.; Büscher, O.; Mütterthies, M.; Pakzad, K.; Wenzel, T.; Woditsch, S. (2015): *Data-Mining – Gesellschaftspolitische und rechtliche Herausforderungen. Fallstudie 2: Data-Mining mit Geodaten. Gutachten für den Deutschen Bundestag, vorgelegt dem Büro für Technikfolgenabschätzung beim Deutschen Bundestag (TAB), Münster, Berlin, 258 S.*

- Gill, Susannah E. et al. (2007): "Adapting cities for climate change: the role of the green infrastructure." *Built environment* 33.1, 115-133.
- Li, Y. (2007): Data-embedded Research. Homepage der Western Michigan University. <https://cs.wmich.edu/~yang/research/dembed> (Zugriff: 21.06.2016).
- Shekar, S. (2014): What is special about mining spatial and spatio-temporal datasets?. – University of Central Florida, Computer Vision Lab, Video Lectures 2014, Orlando. <https://www.youtube.com/watch?v=jJv87Psy4Dk> (Zugriff: 21.07.2016).
- Shekhar, S.; Zhang, P. (2004): Spatial Data Mining: Accomplishments and research Needs.– Vorlesungsskript GIScience 2004, Spatial Computing Research Group, University of Minnesota, Department of Computer Science and Engineering, Minnesota, 64 S. http://www.spatial.cs.umn.edu/paper_ps/giscience.pdf (Zugriff: 21.07.2016).