Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Aaron Braatz

Raumzeitliches Data-Mining in dynamischen Sensornetzen

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Aaron Braatz

Raumzeitliches Data-Mining in dynamischen Sensornetzen

Bachelorarbeit eingereicht im Rahmen des Studiums
im Studiengang Technische Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer : Prof. Dr. Kai v. Luck
Zweitgutachter : Prof. Dr. Tim Tiedemann

Abgegeben am 23.08.2019

Aaron Braatz

Thema der Arbeit

Raumzeitliches Data-Mining in dynamischen Sensorsystemen

Stichworte

Data-Mining, räumliches Data-Mining, raumzeitliches Data-Mining, Maschinelles Lernen, Kriging, LSTM, Neuronale Netze, Crowd Sensing, Dynamische Sensorsysteme

Kurzzusammenfassung

Raumzeitliches Data-Mining findet in fast allen Big-Data-Systemen Anwendung. Zu jedem generierten Datenpunkt werden auch Metadaten gespeichert. Diese enthalten Informationen über den Ort und Zeitpunkt der Generierung. In dieser Arbeit wird der Feinstaubdatensatz von luftdaten.info benutzt. Die Daten werden über ein großes dynamisches Sensorsystem in einem Crowd-Sensing-Kontext erhoben. An diesen dynamischen zeitreihenbasierten Realdaten werden verschiedene raumzeitliche Data-Mining-Verfahren angewendet und evaluiert.

Aaron Braatz

Title of the paper

Space-time data mining in dynamic sensor systems

Keywords

Data Mining, Spatial Data Mining, Spatiotemporal Data Mining, Machine Learning, Kriging, LSTM, Neural Networks, Crowd Sensing, Dynamic Sensor Systems

Abstract

Space-time data mining is used in almost all big data systems. Metadata is also stored for each generated data point. These contain information about the place and time of generation. In this work the particulate matter data set of luftdaten.info is analysed. The data is collected via a large dynamic sensor system in a crowd-sensing context. Different spatiotemporal data mining methods are used and evaluated on these dynamic time series based real data.

Inhaltsverzeichnis

1 Einleitung

Raumzeitliches Data-Mining hat ein großes Anwendungsgebiet. Diverse Themenbereiche haben einen Bezug zu Raum und Zeit. Sei es in der Klimaforschung, Neurowissenschaften oder im Bereich Transport, aber noch in vielen mehr. Jede Thematik mit Bezug zur echten Welt hat eine räumliche und zeitliche Komponente. In Zeiten von „Internet of Things“ und „Industry 4.0“ werden große Mengen multidimensionaler Daten generiert und dazu auch die Raumzeit-Daten (spatio-temporal data, ST) gespeichert.
In der Neurologie werden zu jeder Messung der Gehirnaktivität auch Ort der Aktivität im Gehirn und der Zeitpunkt der Messung dokumentiert. Google speichert zu jeder Internetanfrage den Ort und den Zeitpunkt, an dem die Anfrage gestellt wurde (Spatio-Temporal Data Mining: A Survey of Problems and Methods, 2018 S. 83:2).
Durch immer günstigere Speichermöglichkeiten können diese auch archiviert werden. Leistungsstarke Hardware ermöglicht es, diese Datenmassen zu verarbeiten. Mit diesem Trend geht die Erwartung einher, automatisiert Wissen und Schlüsse für die Zukunft zu sammeln, also aus den Erfahrungen zu lernen.
Auf Grund der Menge und Komplexität der Daten ist es nicht mehr möglich, die Daten händisch auszuwerten. Data-Mining-Verfahren ermöglichen es, Daten zum Großteil automatisiert zu analysieren, Zusammenhänge zu extrahieren und Erkenntnisse zu gewinnen.
Im Zuge der Thematik der zunehmenden Luftverschmutzung und damit auch der Feinstaubbelastung wurde durch das OK Lab Stuttgart ein *Crowd Sensing* Projekt gestartet. Hierbei kann jede Person daran teilhaben, in dem sie eine Messstation aufstellt und die Feinstaubdaten teilt. Dadurch hat sich ein großes dynamisches Sensorsystem mit weltweit etwa 9500 Messstationen gebildet und die Anzahl steigt weiter. Allein im Jahr 2018 wurden nur in Deutschland fast 750 Millionen Messungen mit dem Feinstaubsensor aufgezeichnet, jeweils mit Datum- und Zeitstempel. Diese Daten enthalten potenziell Informationen über die Zusammenhänge oder Entwicklung von Feinstaub.
Da diese Menge an Daten nicht mehr händisch ausgewertet werden kann, ist Gegenstand dieser Arbeit die Untersuchung dieses Datensatzes und Aufbereitung der sensorbasierten Rohdaten aus einem dynamischen Kontext. Weiter soll die Eignung dieses Datensatzes für

raumzeitliche Data-Mining-Verfahren überprüft werden. Dafür werden grundlegende Algorithmen angewendet und evaluiert. Diese Erkenntnisse sollen die Basis bilden für die Anwendung komplexerer Algorithmen.
Die Arbeit ist in 4 Kapitel gegliedert. In Kapitel 2 wird zunächst die Problemstellung analysiert. Hierbei wird auch ein Bezug zu dem aktuellen Stand der Thematik hergestellt. Weiter wird ein Prozess betrachtet, der schrittweise beschreibt, um von den Rohdaten hinzu einem Ergebnis zu gelangen. Daran anschließend werden zwei raumzeitliche Data-Mining-Verfahren vorgestellt. In Kapitel 3 werden die Erkenntnisse aus Kapitel 2 auf den Datensatz angewendet. Ein besonderes Augenmerk liegt dabei auf dem raumzeitlichen Kontext der Daten. Daraufhin werden die Ergebnisse der genutzten Algorithmen ausgewertet und beurteilt. In Kapitel 4 wird abschließend ein Ausblick auf mögliche Weiterentwicklungen und Verbesserungen gegeben.

2 Analyse

2.1 Problemstellung

Mit allen gespeicherten Daten geht die Hoffnung einher, aus ihnen Wissen und relevante Informationen zu gewinnen, um eine gegebene Anwendung zu verbessern.
Bei großen Datenmengen der heutigen Zeit bietet sich Data-Mining (DM) an. Hierbei kommen semiautomatisierte Prozess zum Einsatz, welche den menschlichen Aufwand stark reduzieren. Die gängigen DM-Verfahren gehen dabei allerdings von unabhängigen und gleichmäßig verteilten Datenpunkten aus.
ST-Daten hingegen sind von Natur aus abhängig voneinander, gerade durch ihren Bezug zu Raum und Zeit. Dieser Zusammenhang wird als Autokorrelation bezeichnet. Je näher etwas beieinander ist, desto ähnlicher sind die Eigenschaften. Zudem sind ST-Daten heterogen. Das heißt ein kleiner Ausschnitt ist selten repräsentativ für das Gesamtbild. Zum Beispiel änderte sich die Vegetation in der Nähe von Gewässern, aber auch in Abhängigkeit von der Zeit durch die verschiedenen Jahreszeiten. Daher kann auch nicht wie bei üblichen DM-Verfahren von der gleichmäßigen Verteilung der Daten ausgegangen werden.
Für die Analyse von ST-Daten hat sich daher die Spezialisierung *Spatio-Temporal-Data-Mining* (STDM) gebildet. Hierbei gibt es verschiedene Herangehensweisen:

1. Ein Ort wird als Objekt gesehen und die Messungen über die Zeit ergeben die Attribute.
2. Ein Zeitpunkt wird als Objekt gesehen und die Messungen an den unterschiedlichen Orten ergeben die Attribute.
3. Ereignisse werden als Objekte gesehen und die räumliche und zeitliche Zuordnung ergeben die Attribute.

Ersteres bietet sich für Klima Messungen an, wobei die zeitliche Entwicklung zum Beispiel von Feinstaubdaten beobachtet wird (Discovery of Climate Indices Using Clustering, 2003). Zweiteres wird zum Beispiel in der Neurowissenschaft genutzt, um aktive Gehirnareale nach einer Stimulation zu einem festen Zeitpunkt zu beobachten (Liu, et al., 2018). Letzteres hat unter anderem in der Kriminologie Anwendung. Hierbei werden Straftaten zu bestimmten Orten und Zeitpunkten zugeordnet, um Muster zu erkennen.
In STDM werden unterschiedliche Datentypen betrachtet: Ereignisdaten, welche ein diskretes Ereignis zu gegebenem Ort und Zeitpunkt beschreiben. Bewegungsdaten, wobei die Bewegung im Raum über die Zeit dokumentiert wird. Bei punktbezogenen Daten werden Messungen in einem wandelnden raumzeitlichen Kontext vorgenommen. Rasterdaten

beschreiben Messungen zu einem bestimmten Zeitpunkt und Ort. Ein Teilbereich von STDM ist es, diese Datentypen untereinander zu transformieren, um sie für ein Verfahren nutzbar zu machen.
Die beiden Hauptanwendungen sind allerdings zum einen das Clustering und zum anderen die Vorhersage innerhalb eben genannter Datentypen. Das Clustering unterteilt die Daten in Untergruppen anhand eines Ähnlichkeitsmaßes. Zum Beispiel können Intervalle einer Zeitreihe Ähnlichkeiten zu einer Zeitreihe an einem anderen Ort aufweisen. Innerhalb von Rasterdaten können Bereiche mit ähnlichen Werten gruppiert werden.
Bei der Vorhersage können zukünftige Werte einer Zeitreihe geschätzt werden und innerhalb von Rastern die Punkte zwischen den Messpunkten.
Für diese Arbeit werden insbesondere Rasterdaten betrachtet, da diese typisch für Sensornetze sind. Charakteristisch liegen die Daten in Zeitreihen und in einem räumlichen Raster vor, auch wenn dieses meist nicht geordnet ist.
Dabei liegen die Daten in einer bestimmten Auflösung vor. Landsat-Satelliten nehmen zum Beispiel Messungen von der Erdoberfläche mit einer räumlichen Auflösung von 30 m alle 16 Tage auf. Wird eine höhere Auflösung benötigt, bietet das *Geographische Informations System* (GIS) die Möglichkeiten der Interpolation. Diese genauere Auflösung erhöht den Rechenaufwand für STDM-Verfahren allerdings deutlich. Andererseits kann die Auflösung reduziert werden, um Redundanzen durch die Autokorrelation zu vermeiden.

2.2 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) beschreibt ein iteratives und teils automatisiertes Verfahren zur Extraktion von Wissen aus Datenbeständen. Ein Verfahren wie der KDD-Prozess gewinnt vor allem heutzutage immer mehr an Relevanz. Bei der wachsenden Datenmenge ist ein strukturiertes Vorgehen notwendig, um effizient zu Ergebnissen zu gelangen.
Durch Fayyad et al. wird erstmalig der KDD-Prozess beschrieben als:

> *„KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. “* (Fayyad, et al., 1996 S. 40-41)

Vor dem Beginn des Prozesses wird ein Ziel im Sinne der Anwendung definiert und anhand dessen fünf Schritte durchlaufen, welche in Abbildung 1 dargestellt sind.

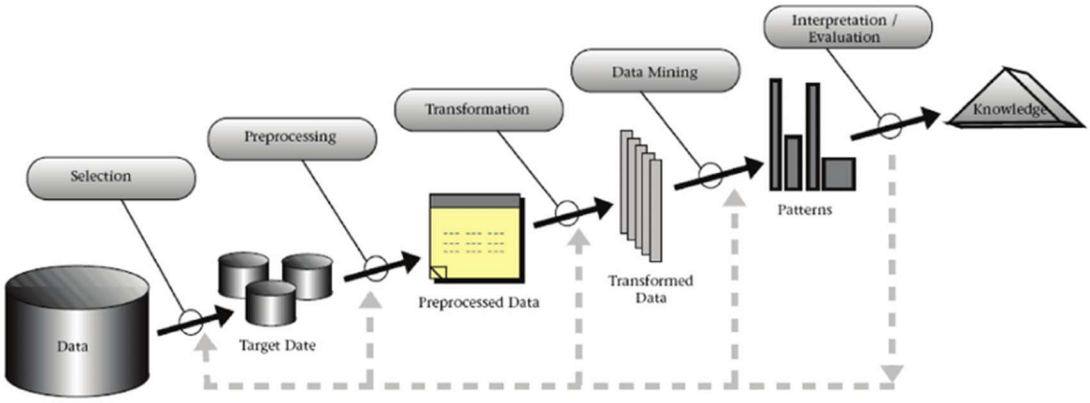

Abbildung 1: KDD-Prozess nach Fayyad [Fayyad et al.1996 Abb. 1]

Zuerst werden in der Datenselektion die Daten ausgewählt oder erhoben. Im zweiten Schritt der Datenvorverarbeitung werden die Daten integriert, konsistent gemacht und gegebenenfalls vervollständigt. In der Datentransformation werden die Daten in ein geeignetes Format überführt, um sie im vierten Schritt einem Data-Mining-Verfahren zu übergeben. Im letzten Schritt wird das Ergebnis interpretiert und evaluiert, ob einer der Schritte angepasst werden muss oder das Ergebnis der Zielsetzung entspricht.

2.2.1 Datenselektion

Die Datenselektion bildet den ersten Schritt des KDD-Prozesses. Hierbei wird die Datenquelle für das weitere Vorgehen definiert. Im Sinne des zuvor festgelegten Ziels muss zuerst ermittelt werden, ob es bereits einen Datenbestand gibt, der das nötige Wissen enthält. Sofern keine passenden Datenbestände existieren, müssen diese während der Datenselektion erstellt werden. Hierzu können zum Beispiel Umfragen oder Messungen genutzt werden. Durch die Erhebung von Daten erhöht sich der Aufwand deutlich. Andernfalls können bereits bestehende Datenbanken genutzt werden. Diese müssen daraufhin gesichtet und nach Möglichkeit mit Domänenwissen bewertet werden (Cleve, et al., 2016 S. 12). So lassen sich aus dem Datensatz sicher Attribute entfernen, welche keinen Nutzen oder gar kontraproduktiv im Sinne der Anwendung sind.
Enthält eine Datenbank nicht alle gewünschten Attribute, ist es wiederum möglich, mehrere Datenbanken zusammenzuführen. Hierbei ist darauf zu achten Redundanzen zu vermeiden, um eine möglichst hohe Informationsdichte zu gewährleisten. Außerdem kann sich bei mehreren genutzten Datenbanken ein größerer Aufwand für die Datenvorverarbeitung ergeben, da bei mehreren Quellen besonders auf die Vereinheitlichung der Daten geachtet werden muss. Eine spezielle Form von Datenbanken sind so genannte Data Warehouse. Diese setzen sich meist aus mehreren Datenbanken zusammen und zeichnen sich durch Themenorientierung, Integrität, Konsistenz und Beständigkeit aus (Witten, et al., 2001 S. 53). Diese Eigenschaften ersparen Aufwand im folgenden Schritt der Datenvorverarbeitung.

Auch haben diese Eigenschaften eine besondere Relevanz für Zeitreihen. Die meisten Data-Mining-Verfahren gehen davon aus, dass die einzelnen Datenpunkte in einem Datensatz unabhängig voneinander sind. Im Falle von Zeitreihen haben die Datenpunkte aber einen gemeinsamen zeitlichen Kontext. Daher sollte schon bei der Datenselektion, insbesondere bei einem Zusammenführen von Datenquellen, auf die zeitlichen Zusammenhänge geachtet werden (Runkler, 2010) (Sartorius, 2019).

2.2.2 Datenvorverarbeitung

Die Datenvorverarbeitung ist der zweite Schritt im KDD-Prozess und dient der Aufbereitung und Steigerung der Qualität der Daten. Dieser Schritt bringt üblicherweise einen hohen Aufwand mit sich, ist aber auch von hoher Wichtigkeit. Je höher die Qualität der Daten ist, desto besser kann die Qualität des Ergebnisses ausfallen. Rohdaten sind häufig fehlerbehaftet, verrauscht, unvollständig, inkonsistent und müssen vorverarbeitet werden (Runkler, 2010 S. 21).
Werden mehrere Datenquellen genutzt, ist es wichtig diese zusammen zu führen und eine gemeinsame Attribuierung festzulegen. Kommen die Datensätze aus unterschiedlichen Abteilungen, ist es möglich, dass Namen jeweils anders vergeben wurden. Daher ist darauf zu achten, ob bei vermeintlich gleichem Attributnamen auch die gleiche Art Daten beschrieben ist oder andersherum ob die gleichen Daten eventuell unterschiedliche Bezeichnungen haben. Bei einer Kategorisierung, ob das Wetter gut oder schlecht ist, sollte zum Beispiel bei dem Attribut Wind gleich definiert sein, bis zu welcher Stärke das Wetter noch gut ist. Sind die gleichen Daten mehrfach vorhanden, sollten sie aus Gründen der Redundanz verworfen werden. Auch ist es wichtig, ob die Daten mit einer vergleichbaren Messmethode erhoben wurden und die Daten im gleichen Format gespeichert sind. Um diese Entscheidungen korrekt zu treffen, wird Domänenwissen benötigt. Bei Zeitreihen kann es sinnvoll sein, den Betrachtungsbereich auf die vorhandenen Daten einzuschränken. Dementsprechend wird für den frühesten Datenpunkt der Zeitpunkt null definiert und alle weiteren Datenpunkte werden entsprechend angepasst. Durch diesen Schritt wird die Bedeutung der Zeit auf den betrachteten Bereich reduziert, da meistens nicht die Uhrzeit oder ein Datum für die Analyse relevant ist, sondern viel mehr die zeitliche Kausalität.
Um die Problematik mit fehlenden oder fehlerbehafteten Daten anzugehen, gibt es mehrere Strategien. Zum Beispiel lassen sich Fehler durch Rauschen und Ausreißer in Zeitreihen unter anderem durch Filter reduzieren. Da die Daten in dieser Thesis Zeitreihen entsprechen, wird im Folgenden besonders auf Strategien in diesem Kontext eingegangen.
Die Problematiken mit Rohdaten lassen sich grob in fehlende und fehlerhafte Daten einteilen.

Fehlende Daten können zum Beispiel durch Versagen der Messgeräte oder durch nicht gegebene Antworten in einer Umfrage auftreten. Grundsätzlich geben die meisten Data-Mining-Verfahren fehlenden Werten keine besondere Bedeutung und kein weiteres Handeln ist nötig. Je nach Anwendungskontext kann ein fehlender Wert aber einen besonderen

Informationsgehalt haben. Wurden bei einer Umfrage auf eine Frage zum Einkommen von mehreren einer bestimmten Personengruppe keine Antwort gegeben, hat diese Frage möglicherweise einen Bezug zu der Personengruppe. In diesem Fall sollten diese Werte mit einem Attribut „verweigert“ bezeichnet werden (Ester, et al., 2000 S. 3). Bei Zeitreihen sollten die fehlenden Werte mit einem geeigneten Verfahren approximiert werden. Mögliche Methoden dafür können auch bei der Ersetzung von Ausreißern angewandt werden und sind im folgenden Abschnitt genannt.

Bei fehlerhaften Daten gibt es zwei unterschiedliche Arten von Fehlern in Daten, beziehungsweise Messwerten. Zum einen gibt es systematische Fehler. Hierbei treten Fehler durch fehlerhafte Kalibrierung der Messgeräte, falsche Skalierung der Daten oder Drifteffekte auf. Diese Fehler lassen sich grundsätzlich vollständig ausgleichen, wenn die Systematik bekannt ist.
Die zweite Fehlerart sind zufällige Fehler. Zu denen gehören zum Beispiel Mess- und Übertragungsfehler. Diese können häufig als additives Rauschen modelliert und mit Hilfe von Filterung korrigiert werden. Allerdings gilt das nicht für alle zufälligen Fehler. Insbesondere Ausreißer lassen sich nicht so modellieren und können von besonderer Relevanz sein. Einerseits können sie durch Fehler in der Erfassung und Verarbeitung verursacht werden. Wenn die Daten teilweise durch Menschen erfasst werden, kann es zu Fehlern wie Zahlendrehern kommen. Auch wenn die Daten zwischen Systemen übertragen und unterschiedliche Datenformate benutzt werden, können zufällige Fehler auftreten. Andererseits können ungewöhnliche Daten auch einen realen Ursprung haben. Im Beispiel der Feinstaubanalyse wäre es fatal, wenn starke Ausschläge grundsätzlich als Messfehler oder Ausreißer deklariert werden, obwohl tatsächlich eine hohe Feinstaubbelastung vorliegt. Hierfür ist es hilfreich die Daten zu visualisieren, um eine Einschätzung zur Ursache eines Ausreißers geben zu können. Dazu müssen die Ausreißer erst einmal erkannt werden. Zum einen kann überprüft werden, ob sich der Wert in einem gültigen Wertebereich befindet. Ist der Wert außerhalb des Wertebereichs des Messgeräts lässt sich über eine passende Regel ein Ausreißer erkennen.
Unterscheidet sich ein Ausreißer von allen anderen Daten kann die Abweichung über Methoden der Statistik erkannt werden. Zum Beispiel werden mit der Sigma-Regel Ausreißer erkannt, die um mehr als die Standardabweichung vom Mittelwert der gesamten Daten abweicht. Diese Regel lässt sich auch mit vielfachen der Standardabweichung anwenden, um eine größere Abweichung als „normal“ zu definieren und nur Daten mit besonders starker Abweichung als Ausreißer zu deklarieren (Sartorius, 2019 S. 224). So werden mit der 2-Sigma-Regel Daten mit einem Unterschied von mehr als zwei Standardabweichungen vom Mittelwert als Ausreißer bestimmt.
Nach der 2-Sigma-Regel ist der Datenpunkt x_k ein Ausreißer, wenn:

$$\left|\frac{x_k - \bar{x}}{s}\right| > 2$$

Wobei $\bar{x}$ der Mittelwert und s die Standardabweichung ist.

Ausreißer, die lokal abweichen aber in dem Wertebereich der Daten bleiben, lassen sich mit der Methode so nicht erkennen. Ist ein Datenpunkt in einem allgemeinen Cosinus Signal, welcher sich nahe einem Minimum befindet, unerwarteterweise auf eins, dann lässt sich der Ausreißer nicht mit der Sigma-Regel erkennen.
Solche Ausreißer lassen sich in Relation zum näheren Umfeld des Datenpunktes erkennen, wie dem vorhergehenden oder nachfolgendem Wert.

Für die Bearbeitung von Ausreißern gibt es verschiedene Möglichkeiten:
Der Ausreißer wird als solches markiert: Dadurch wird der Datensatz nicht verändert, aber die Information über den Ausreißer kann in folgenden Verarbeitungsschritten berücksichtigt werden. Eine Möglichkeit die Information zu speichern, ist eine Markierungsmatrix mit der gleichen Größe wie der Datensatz. Dabei wird für jeden Wert im Datensatz korrespondierend ein True für Ausreißer und ansonsten False gespeichert. Diese Variante ist durch die zusätzliche Speicherung der Markierungsmatrix besonders ineffizient für große Datensätze.
Der Ausreißer wird entfernt: Um das zu erreichen, kann der einzelne Wert mit einem besonders definierten Symbol bezeichnet werden, wie zum Beispiel NaN (Not a Number) oder einem fest definierten Wert, der nicht in dem Spektrum des Wertebereichs vorkommt. In einem Bereich mit nur positiven Werten könnte zum Beispiel „-1“ gewählt werden.
Der ganze Datenpunkt mit Ausreißer wird entfernt: Diese Methode wird häufig angewandt, hat aber vor allem in stark fehlerbehafteten Datensätzen die Folge, dass viele Daten verloren gehen. Insbesondere ist diese Möglichkeit für Zeitreihen kontraproduktiv, da der zeitliche Kontext verletzt wird.
Die ganze Datenreihe, die den Ausreißer enthält, wird entfernt: Diese Methode ist nur empfehlenswert, wenn die Datenreihe viele Ausreißer enthält. Allerdings wurde die Datenreihe bei der Datenselektion vermutlich bewusst ausgewählt und es sollte möglichst eine alternative Datenquelle gesucht werden.
Der Ausreißer wird korrigiert: Dafür stehen mehrere Strategien zur Verfügung:

1. Ersetzen durch den Maximal- / Minimalwert
2. Ersetzen durch den globalen Mittelwert
3. Ersetzen durch den nächsten Nachbarn, der kein Ausreißer ist
4. Interpolation bei Zeitreihen
5. nichtlineare Interpolation, zum Beispiel mit Splines
6. Filterung
7. modelbasierte Ergänzung, zum Beispiel durch Regression

(Runkler, 2010 S. 23) (Grzymala-Busse, et al., 2010 S. 35-43)

2.2.3 Datentransformation

Die Datentransformation ist der dritte Schritt des KDD-Prozesses und dient vor allem dazu den Datensatz in ein Format zu überführen, welches für das Data-Mining-Verfahren nutzbar

ist. Zudem kann noch eine genauere Attribut-Selektion vorgenommen werden. Diese ist auch bekannt als *Feature-Selection*. Auch wenn die meisten Lernverfahren implizit ausgelegt sind, relevante Attribute zu erkennen, haben Experimente mit dem Entscheidungsbaum-Lernsystem „C4.5“ gezeigt, dass die Klassifizierungsleistung durch Hinzufügen willkürlicher Attribute verschlechtert wird (je nach Situation 5-10 %) (Witten, et al., 2001 S. 252). Der Grund ist, dass an einem Punkt in dem Verfahren auch die willkürlichen Attribute zur Ergebnisfindung herangezogen werden. Für die Entscheidung, welche Attribute die höchste Relevanz für das Ergebnis haben, ist Domänenwissen nötig. Allerdings gibt es auch algorithmische Ansätze für die Auswahl. Hierbei wird zwischen der verfahrensunabhängigen und -abhängigen Auswahl unterschieden. Erstere wird als Filter-Methode bezeichnet, da sie die gesamte Attributmenge filtert, um eine günstige Teilmenge zu erzeugen. Hierbei wird versucht, ein Maß an Relevanz eines Attributs hinsichtlich des Ergebnisses zu erstellen. Leider gibt es dafür kein allgemeingültiges Maß und so haben die Algorithmen unterschiedliche Ansätze und sind je nach Aufgabenkontext zu wählen. Zweiteres wird als Wrapper-Methode bezeichnet, da der Lernalgorithmus zur Entscheidungsfindung mit einbezogen wird. Hierbei wird die Attributteilmenge an der Leistung des Lernverfahren gemessen. Die Leistung wird meist mittels Kreuzvalidierung geschätzt, aber auch andere Evaluierungsmethoden können genutzt werden. Grundsätzlich ist darauf zu achten, dass durch die Auswahl der Attributteilmenge nicht nur diejenigen ausgewählt werden, die beschreibend für den bisherigen Datensatz sind, sondern auch für kommende Daten verallgemeinert werden können, also Überanpassung (engl. Overfitting) vorgebeugt wird (Witten, et al., 2001 S. 254).

Je nach verwendetem Data-Mining-Verfahren müssen die Daten in numerischer oder kategorischer Form übergeben werden. Liegen die Daten in numerischer Form vor, aber das Verfahren verarbeitet kategorische Daten, müssen diese diskretisiert werden. Hierfür kann der Wertebereich in Intervalle gleicher Größe geteilt werden oder in Intervalle mit gleicher Häufigkeit eines Attributs. Komplexere Verfahren können die Intervalle auch an eventuell bekannte Klassenzugehörigkeiten anpassen. Wird eine numerische Form von dem DM-Verfahren erwartet, werden die Attribute mittels „One-Hot-Encoding“ in den numerischen Bereich transformiert. Hierbei wird ein Vektor mit einer Größe gleich der Anzahl unterschiedlicher Kategorien eines Attributs gebildet. Jede Stelle ist stellvertretend für genau eine Kategorie und wird mit einer „1“ markiert, sofern die Kategorie auf den Datenpunkt zu trifft (Géron, 2018 S. 64).
Um Attributen mit nummerisch höheren Werten nicht implizit eine stärkere Gewichtung zu geben, ist es üblich, den Wertebereich zu normalisieren. Hierbei werden die Attribute auf einen Wertebereich von 0…1 beziehungsweise -1…1 abgebildet.
Normalisierung auf den Bereich 0…1:

$$y_k = \frac{x_k - x_{min}}{x_{max} - x_{min}}$$

Wobei die neuen Intervallgrenzen durch das Maximum und Minimum des Attributs definiert sind. Hierbei ist weiter zu beachten, dass das gleiche Maximum und Minimum auch für

zukünftige Daten zur Normalisierung genutzt werden, um die Relation beizubehalten (Chollet, 2018 S. 139).
Bei Zeitreihendaten ist in diesem Arbeitsschritt darauf zu achten, dass in allen Datenreihen die Messungen in gleichen Intervallen vorgenommen wurden. Ist das nicht der Fall, müssen die Datenreihen angepasst werden und ggf. Datenpunkte approximiert werden. In dem gleichen Zuge lässt sich auch die Frequenz der Daten anpassen, um eventuell die Datenmenge für bessere Rechenzeiten in folgenden Schritten zu reduzieren.

2.2.4 Data-Mining

Das Data-Mining ist der vierte Schritt des KDD-Prozesses und beinhaltet die eigentliche Wissensgewinnung. Mit Algorithmen der Statistik und des Maschinellen Lernens (engl. *Machine Learning* ML) wird das Wissen aus den zuvor aufbereiteten Daten extrahiert. Diese Algorithmen können in vier Hauptdisziplinen unterteilt werden: Klassifikation, Clustering, Gewinnung von Assoziationsregeln und Generalisierung. Mittlerweile wird aus dem Bereich des ML auch die Regression dazu gezählt. Einige der Algorithmen können schon in der Datenvorverarbeitung und Datentransformation angewendet werden, um fehlende oder fehlerhafte Daten mittels Regression zu ermitteln. Außerdem kann mit Algorithmen der Generalisierung die Anzahl der Attribute reduziert werden.
Vor der Auswahl eines Algorithmus wird üblicherweise ein einfaches Baseline-Verfahren angewendet, um möglichst kostengünstig die Brauchbarkeit des Datensatzes zu ermitteln. Werden mit diesem Verfahren keine brauchbaren Ergebnisse ermittelt, ist entweder der Informationsgehalt in dem Datensatz zu gering oder zu komplex für das einfache Verfahren. Im zweiten Fall gibt es noch die Möglichkeit, ein komplexeres Verfahren anzuwenden. Andernfalls ist der Datensatz vermutlich nicht für das Ziel der Anwendung geeignet.
Sollte das Baseline-Verfahren erfolgversprechende Ergebnisse liefern, kann ein spezifischerer Algorithmus gewählt werden. Welcher Algorithmus verwendet wird, hängt von dem Ziel der Anwendung und der Art der vorliegenden Daten ab. Dieser kann in der Evaluation auch mit dem Baseline-Verfahren verglichen werden (Giudici, 2010 S. 649). Je nach Ziel kann die Menge der Algorithmen anhand einer der obengenannten Disziplin eingegrenzt werden. Und wie bereits im vorhergehenden Schritt des KDD-Prozesses beschrieben, kann ein bestimmtes Datenformat Voraussetzung für das Data-Mining-Verfahren sein. So kann ein passender Algorithmus identifiziert werden.
Bei SDTM-Verfahren wird zusätzlich zu der zeitlichen Dimension auch die räumliche berücksichtigt. Einer der wichtigsten Zusammenhänge ist die räumliche Autokorrelation und wird auch als erstes Gesetz der Geografie bezeichnet.

> *"Everything is related to everything else but nearby things are more related than distant things"* (Tobler, et al., 1970 S. 236)

2.2.5 Evaluation

Die Evaluation ist der fünfte und letzte Schritt in dem KDD-Prozess. Hierbei werden die Ergebnisse, beziehungsweise das Modell, in einer geeigneten Form präsentiert und bewertet. Für die Präsentation der Ergebnisse müssen diese gegebenenfalls noch visuell oder textuell aufbereitet werden.
Die Bewertung findet in Hinsicht auf das Ziel der Anwendung statt und wird von einem Experten der Domäne durchgeführt. Hierbei wird nach den oben genannten Zielen des KDD-Prozesses bewertet: Ist das Ergebnis gültig, neuartig, nützlich und verständlich (Fayyad, et al., 1996 S. 40-41).
Zum einen können die Ergebnisse subjektiv durch den Kunden oder Anwender bewertet werden oder zum anderen durch objektive Metriken, zum Beispiel die Fehlerrate, welche vor allem beim überwachten Lernen angewendet wird. Hierbei wird das Verhältnis der fehlerhaften Ergebnisse zu der Gesamtzahl der Ergebnisse gemessen (Zhang, 2010 S. 425).
Aus der Perspektive der Nützlichkeit soll das Modell auch auf zukünftige Daten anwendbar sein. Die sogenannte Vorhersagekraft kann daran geschätzt werden, wie gut sich das Modell auf den Testdatensatz verallgemeinern lässt. Dieser ist dem Modell während des Trainings nicht bekannt und kann somit ein Indiz dafür geben, wie das Modell mit neuen Daten umgeht. Insbesondere bei kleinen Datensätzen ist eine Kreuzvalidierung empfehlenswert.
Ist das Ergebnis noch nicht zufriedenstellend, wird eine neue Iteration des KDD-Prozesses initialisiert. Dabei kann an jedem Schritt des Prozesses eingestiegen werden und Änderungen vorgenommen werden. Das kann dazu führen, dass ein neuer Datensatz gesucht, die Vorverarbeitung angepasst oder ein anderes Data-Mining-Verfahren ausgewählt werden muss. Da die einzelnen Schritte abhängig voneinander sind, sollten Änderungen einzeln vorgenommen werden, damit die Auswirkungen nachvollziehbar sind (Ester, et al., 2000).

2.3 Zeitreihen Analyse

Die wesentlichen Analysemöglichkeiten von Zeitreihen in einem STDM-Kontext sind einmal die Gruppierung (*Clustering*) von Zeitreihen anhand ihres Verhaltens und andererseits die Vorhersage weiterer Datenpunkte in einer Zeitreihe. Für das Clustering der Zeitreihen gibt es verschiedene Ansätze. Die traditionellen Algorithmen wie *k-means* (Mezer, et al., 2009), *hierachical clustering* (Goutte, et al., 1999), *shared nearest neighbor clustering* (Discovery of Climate Indices Using Clustering, 2003) und *normalized-cut spectral clustering* (van den Heuvel, et al., 2008) können verwendet werden, allerdings ist dabei nicht der räumliche Zusammenhang gewährleistet. In (Bellec, et al., 2006) (Heller, et al., 2006) (Lu, et al., 2003) (Blumensath, et al., 2013) (Craddock, et al., 2012) werden Algorithmen beschrieben, die die räumlichen Beziehungen berücksichtigen.
Für die Vorhersage von Zeitreihen sind Neuronale Netze mittlerweile gängige Praxis (Spatio-Temporal Data Mining: A Survey of Problems and Methods, 2018 S. 16f.).

2.3.1 Deep Learning Ansatz

Deep Learning (DL) ist ein Teilbereich des ML. Hierbei werden komplexe Zusammenhänge mit ähnlichen Prozessen wie im menschlichen Gehirn ermittelt. In künstlichen neuronalen Netzen (KNN) werden die Daten in mehrschichtigen und verknüpften Perzeptronen verarbeitet. Die erste Schicht nimmt die Daten entgegen und wird als Input-Layer bezeichnet. Darauf können beliebig viele Hidden-Layer, also die Schichten mit den Neuronen folgen. KNN abstrahieren selbständig Informationen aus den zur Verfügung gestellten Daten. Ein einzelnes Neuron in einem KNN besteht aus Gewichten, welche jedem Eingangswert der „Vor-Neuronen" eine bestimmte Priorisierung und ein Bias zuweist. Diese Netzparameter werden während der Trainingsphase optimiert und somit die Abstraktion der Informationen verbessert. Die Gewichte und Bias werden in Matrizen abgebildet, welche je nach Komplexität des KNN sehr groß werden können (Manaswi, 2018 S. 46).
Das Training eines KNN kann je nach ML-Ansatz unterschiedlich ausfallen. Zu unterscheiden ist hierbei zwischen drei Lernmethoden. Das *Überwachte Lernen* (engl. *supervised learning*) lernt anhand von bereits klassifizierten Daten, also gibt es mit den Trainingsdaten das erwartete Ergebnis (Label). Bei dem *unüberwachten Lernen* (engl. *unsupervised learning*) werden Informationen rein aus den Daten gewonnen, zum Beispiel beim Clustering von Daten. Das *verstärkende Lernen* (engl. *reinforcement learning*) trainiert durch ein Belohnungssignal. Im Gegensatz zu dem überwachten Lernen ist das Belohnungssignal nicht durch die Trainingsdaten bereitgestellt, sondern wird durch die Anwendungsumgebung und der Entscheidung des KNN generiert (Géron, 2018 S. 8ff.).
In dieser Arbeit wird das überwachte Lernen betrachtet, da zu den Daten die Label (die zukünftigen Daten) bekannt sind. In diesem Fall ermittelt während des Trainings eine Verlustfunktion die Differenz zwischen dem Ergebnis des KNN und dem Label. Mit dem Backpropagation-Algorithmus werden die Netzparameter angepasst, sodass die Distanz gegen null konvergiert. Hierbei wird das Gradientenverfahren genutzt, um den Rechenaufwand zu minimieren. Durch die Anpassung sogenannter Hyperparameter ist es möglich, das Training effektiver zu gestalten und das KNN genauer an den Anwendungsfall anzupassen (Chollet, 2018 S. 30).

Vorhersage mit RNN

Für die Verarbeitung von Sequenzen haben sich die Rekurrenten neuronalen Netze (RNN) etabliert. Dementsprechend eignen sich RNN vor allem für die Verarbeitung von Texten, Videos, Sprache, Geninformationen und Zeitreihen. RNN können für Klassifizierung, Clustering und insbesondere Vorhersage von Texten und Zeitreihen genutzt werden. Die Besonderheit eines RNN gegenüber eines KNN ist eine Feedback-Schleife, welche die vorherigen Ergebnisse speichert. Dadurch werden für die Verarbeitung neuer Daten auch die Informationen über vorhergegangene Daten berücksichtigt. Dadurch sind (tiefe) RNN deutlich komplexer und haben mehr trainierbare Netzparameter und damit auch mehr Gradienten während des Trainings. Durch die große Anzahl an Gradienten über die

Zeitschritte kann es in RNN zu zwei Problematiken kommen. Sind die Gradienten kleiner Eins, kommt es durch die Multiplikation in dem Gradienten-Verfahren zu einer exponentiellen Verringerung der Gradienten (engl. V*anishing*) oder bei Gradienten größer Eins zur exponentiellen Verstärkung (engl. *Exploding*). Dadurch wird das Training von RNN instabil und langsam.
Um dieser Problematik entgegen zu wirken, gibt es eine spezielle Form der RNN: *Long-Short-Term-Memory-Networks* (LSTM). Hierbei werden durch zusätzliche Gatter in der Feedback-Schleife extreme Gradienten vermieden (Huang, et al., 2019 S. 32).

2.4 Räumliches Data-Mining

Das räumliche Data-Mining (engl. *Spatial-Data-Mining* SDM) lässt sich vor allem in die beiden Bereiche Clustering und Vorhersage von räumlichen Daten aufteilen. Dabei ist die Vorhersage eher die Interpolation im Raum.
Das Clustering kann hier mit unterschiedlichem Schwerpunkt durchgeführt werden. Zum einen können die Daten in ihrer Verteilung im Raum gruppiert werden. Hierbei werden die Datenpunkte durch ihre minimale Distanz zum Cluster-Zentrum einem Cluster zugeordnet. Typische Algorithmen für diesen Ansatz sind *k-means*, der *EM-Algorithmus*, *CLIQUE*, *BIRCH* und *CLARANS* (Celik, et al., 2008). Einen hierarchischen Ansatz bietet *Chameleon* (Karypis, et al., 1999). Ein weiterer Ansatz ist das Clustering über die Dichte der Datenpunkte. Diesen Ansatz verfolgen *DB-Scan* (Ester, et al., 1996) und *shared nearest neighbors* (Jarvis, et al., 1973).
In der räumlichen Interpolation werden Daten für einen Ort anhand der umliegenden Messungen geschätzt. Auch in dieser Disziplin gibt es verschiedene Ansätze: *Spatial auto-regressive models* (SAR) (Kelejian, et al., 1999), *geographically weighted regression models* (GWR) (Brunsdon, et al., 1996), Modelle mit *Markow random field* (Schroder, et al., 1998) und *Kriging* (Oliver, et al., 1990). In Bezug auf Geodaten ist der verbreitetste Algorithmus in der Geostatistik das Kriging (Bhattacharjee, et al., 2014). Im Gegensatz zu anderen Algorithmen werden beim Kriging auch Redundanzen in die Schätzung mit einbezogen (Gau, 2010 S. 34).
Daher wird Kriging in dieser Arbeit betrachtet und im Folgenden noch genauer erläutert.

2.4.1 Geostatistische Schätzverfahren

Die geostatistische Schätzung ist ein Teilbereich der Geostatistik und nimmt Bezug auf die Theorie der regionalisierten Variablen (Gau, 2010 S. 20). Das Ziel geostatistischer Schätzverfahren ist es, die Schätzung der natürlich vorkommenden Zufallsprozesse anhand von punktuell vorhandenen Informationen. So können zum Beispiel von einzelnen Bohrproben Schlüsse auf die dazwischen liegenden Sedimente gezogen werden.
Der Ablauf einer geostatischen Schätzung hat Parallelen zu dem KDD-Prozess und besteht aus mehreren Schritten, über die ein Schätzmodell entwickelt wird. Der erste Schritt ist die

Planung und Durchführung der Erkundungskampagne. Ähnlich wie bei dem KDD-Prozess werden hier die Ziele und der Verwendungszweck des Modells definiert. Im zweiten Schritt der Datenerhebung werden an zuvor bestimmten Orten die Messungen vorgenommen. Als nächstes folgt die Zusammenfassung von Datenkollektiven. Dabei werden die Ergebnisse der Messungen in thematische und für die weitere Modellierung vorteilhafte Gruppen unterteilt. In den nächsten beiden Schritten wird für jedes Kollektiv erst ein experimentelles und ein darauf aufsetzendes theoretisches Variogramm erstellt. Durch die Variogramme wird die schon in Teilabschnitt 2.2.4 erwähnte Autokorrelationsstruktur ermittelt. Anhand dieser Struktur wird dann die eigentliche geostatistische Schätzung durchgeführt. Dies wird üblicherweise als Kriging bezeichnet, in Gedenken an D. G. Krige. Es gibt verschiedene Formen des Kriging, wobei immer gewichtete Mittelwerte ermittelt werden. Als letztes folgt die Visualisierung. Gegebenenfalls können nachträglich noch Modellparameter angepasst werden (Gau, 2010 S. 27). Die drei Schritte, Ermittlung des experimentellen, sowie des theoretischen Variogramms und des Krigings werden im Folgenden noch näher erläutert.

Experimentelle Variographie

In der experimentellen Variographie wird ein Variogramm anhand der bekannten Messungen erstellt. Hierbei wird nach der Theorie der regionalisierten Variablen eine Autokorrelationsstruktur γ unterstellt. Diese Struktur wird durch ein mathematisches Modell geschätzt. Hierbei wird die Varianz der Zufallswerte z zwischen den Messpunkten x innerhalb einer festgelegten Entfernung h (*Lag*) bestimmt und durch die Anzahl dieser Wertepaare n dividiert. Die geschätzten entfernungsbezogenen Autokorrelationen und die Entfernung werden zusammen in dem experimentellen bzw. empirischen Variogramm dargestellt (Oliver, et al., 2015 S. 12).

Hierbei ist üblicherweise zu erkennen, dass bei kleinerer Entfernung die Unähnlichkeit niedrig ist (kleine $\gamma^*_{(h)}$) und bei großen Entfernungen die Unähnlichkeit einem Grenzwert

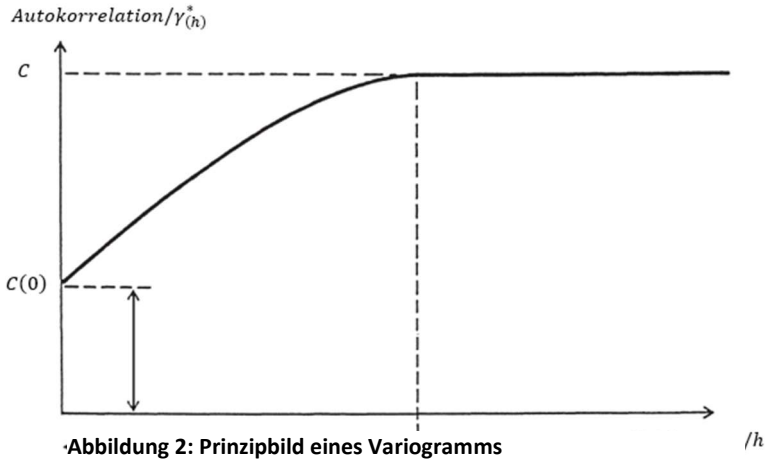

Abbildung 2: Prinzipbild eines Variogramms [Heinrich1994 Abb. 4.7.4 angepasst]

C annähert. Dieser wird als „sill“ bezeichnet und die Entfernung, bei der das sill erreicht wird, wird als range a bezeichnet, wie in Abbildung 2 beschrieben. Üblicherweise ist bei Realdaten

der Ordinatenschnittpunkt nicht gleich dem Ursprung. Die Differenz wird als „nugget-effect“ bezeichnet und folgt meist aus Messfehlern oder einer Mikrovariabilität, welche unterhalb des Probenabstands liegt (Heinrich, 1994 S. 151).

Theoretische Variographie

In der theoretischen Variographie wird ein theoretisches Modell gewählt, das möglichst genau dem experimentellen Variogramm entspricht. Dadurch sollen auch Variogrammwerte relativ zu der Entfernung bestimmt werden, die nicht mit den Messungen abgedeckt sind. Diese werden für die Schätzung der räumlichen Struktur und daraus resultierend auch der einzelnen Zufallswerte benötigt. Das Variogrammmodell wird durch eine mathematische Funktion abgebildet. Damit die Kriging-Gleichung lösbar bleibt und nur positive Varianzen ermittelt werden, muss das Modell positiv-semidefinit sein (Heinrich, 1994 S. S.154). Grundsätzlich wird zwischen transitiven und intransitiven Modellen unterschieden. Das transitive Modell setzt die Stationarität zweiter Ordnung um. Das bedeutet unteranderem, dass der Erwartungswert der Zufallsvariablen an jedem Ort dem Mittelwert der vorhandenen Daten entspricht und das eine Kovarianz zwischen allen Zufallsvariablen existiert. Das intransitive Modell setzt lediglich die intrinsische Hypothese um. Diese besagt, dass der Erwartungswert der Differenz zweier Zufallsvariablen gleich null ist und bei jeder Entfernung zwischen den zwei Punkten die Varianz endlich ist (Gau, 2010 S. 30).
Am häufigsten werden das sphärische, exponentielle und Gauß'sche Modell genutzt, welche alle dem transitiven Ansatz unterzuordnen sind.
Für komplexere Variogramme lassen sich Modelle auch kombinieren.

Kriging

Kriging bezeichnet den eigentlichen geostatistischen Schätzprozess. Hierbei wird mit Hilfe der im Variogramm beschriebenen Autokorrelationsstruktur und den bekannten Datenpunkten, Punktschätzungen gemacht. Kriging ist hierbei nur der Name für eine Menge an Methoden, welche grundsätzlich bei der Schätzung gewichtete Mittelwerte nutzen und versuchen, die Schätzvarianz zu minimieren. Ein Krigingschätzer ist ein exakter Interpolator. Wird ein Punkt geschätzt, der bereits bekannt ist, wird auch der bekannte Wert ermittelt und die Schätzvarianz ist gleich null. Durch die im Kriging abgebildete räumliche Abhängigkeit werden einerseits naheliegendere Messpunkte höher gewichtet, dabei wird aber auch auf Redundanz in den Informationen geachtet. Liegen zwei Messpunkte nahe nebeneinander fallen diese trotzdem nicht doppelt ins Gewicht. Außerdem erhalten Messpunkte, die aus der Perspektive des zu schätzenden Punktes hinter einem anderen Messpunkt liegen, auch ein niedrigeres Gewicht (*screen-effect*) (Heinrich, 1994 S. 158).

Im Weiteren wird vor allem auf das *Ordinary Kriging* (nach (MATHERON, G. (1963): Principles of geostatistics. In: Economic Geology, 58: 1246-1266) vgl. (Schroder, et al., 1998 S. 156)) eingegangen. Hierbei handelt es sich um ein stationäres, lineares und univariantes Kriging. Durch die Robustheit wird dieses Schätzverfahren auch als *„anchor algorithm of geostatistics"* bezeichnet. Dadurch, dass die Gewichte der Mittelwerte hinsichtlich der Varianz der Schätzwerte optimiert werden und die Differenz zwischen wahren und geschätzten Werten im Mittel null ist (Unverzerrtheit), wird ein solcher Schätzer als BLUE bezeichnet. BLUE ist in der traditionellen Schätzstatistik die Abkürzung für „best linear unbiased estimator" (Oliver, et al., 2015 S. 43). Zusätzlich zu den geschätzten Werten lässt sich auch die Schätzvarianz, auch Kriging-Varianz, berechnen. Diese Varianz bietet ein Maß der Sicherheit zu der Schätzung. Ist die Kriging-Varianz größer als der Wertebereich der Daten, sind die Ergebnisse nicht aussagekräftig und das Kriging Modell muss angepasst werden.

2.5 Zielsetzung

Die Thematik der Umweltbelastung und in dem Zuge auch der Feinstaubbelastung findet immer mehr Aufmerksamkeit in der Bevölkerung. Durch die Sensorstationen von luftdaten.info haben auch Privatpersonen die Möglichkeit, direkt vor der eigenen Haustür die Belastung zu messen. Das Sammeln von Daten durch viele Individuen wird auch als *Crowd Sensing* bezeichnet. Dieses verteilte Sammeln von Daten wird als dynamisches Sensorsystem betrachtet.
In dieser Arbeit sollen zunächst die Daten unter Anwendung des KDD-Prozesses und mit den bereits vorgestellten Algorithmen gesichtet und visualisiert werden, um ggf. Problemstellungen aufzudecken und passend zu behandeln. Hierbei wird insbesondere die Art der Daten, als Realdaten, Zeitreihen und mit räumlicher Korrelation, berücksichtigt.
Des Weiteren sollen STDM-Verfahren an den Daten erprobt werden und die Anwendbarkeit überprüft werden. Hierfür wird einerseits ein LSTM-Netz entwickelt, um zu überprüfen, ob mit den Daten Prognosen gemacht werden können. Aus dem Bereich des SDM wird das geostatistische Schätzverfahren mit dem Kriging-Algorithmus erprobt. Ein besonderes Augenmerk wird auf die wandelnden Messorte gelegt, da jeder Nutzer die Sensorstation zu einem beliebigen Zeitpunkt und an einer beliebigen Position aufstellen kann. Aufgrund der großen Menge an Daten und der daraus resultierenden rechnerischen Komplexität wird sich diese Arbeit auf die Messungen im Bereich Stuttgart über das Jahr 2018 begrenzen.

3 Experimentelle Umsetzung

In diesem Kapitel werden am Beispiel des Datensatzes von luftdaten.info einige STDM-Verfahren angewendet. Grundsätzlich orientiert sich der Aufbau an dem KDD-Prozess. In Abschnitt 3.1 wird kurz auf die verwendete Software eingegangen. Darauffolgend wird in Abschnitt 3.2 der Datensatz beschrieben und auf Probleme analysiert. In Abschnitt 3.3 werden diese Probleme dann in der Datenvorverarbeitung behandelt. Im Abschnitt 3.4 werden die Daten aus der zeitlichen Perspektive mittels eines LSTM-Netzes analysiert. Darauf folgt das geostatistische Schätzverfahren aus dem SDM. Abschließend wird noch ein Fazit über die Daten und die angewendeten Algorithmen gezogen.

3.1 Verwendete Software-Toolchain

Für die Umsetzung dieser Arbeit wurden unterschiedliche Softwareprodukte genutzt. Grundsätzlich wurde für die Entwicklung und Implementierung des geostatistischen Schätzverfahrens die Skriptsprache Python in der Version 3.6 verwendet. Für kleinere Analysen wurde MS Excel genutzt. Für die Visualisierung der Daten wurde die Bibliothek Matplotlib genutzt. Die Daten wurden mit Apache Parquet in einem spaltenorientierten Format gespeichert. Für die Entwicklung des LSTM-Netzes wurde das Framework Keras mit Tensorflow verwendet. Durch dieses Framework wird die Entwicklung neuronaler Netze stark abstrahiert und dadurch vereinfacht.

3.2 Datensatz

Citizen Science (Bürgerwissenschaft) ist ein Weg für jeden Bürger, selbst als Laie an wissenschaftlichen Projekten teilzunehmen oder komplett eigenständig aufzubauen. Das *OK Lab Stuttgart* hat sich einem solchen Projekt gewidmet und bietet jedem die Möglichkeit, ein günstiges Feinstaubmessgerät zu erwerben und die eigenen Messungen zu teilen. Mittlerweile gibt es weltweit fast 9500 dieser Sensorstationen. Davon sind in etwa die Hälfte in Deutschland installiert. Für einen Preis zwischen 40-70 € lassen sich die Bauteile erwerben, die daraufhin selbstständig zusammengebaut werden können. Die Station besteht aus einem ESP8266-Microcontroller und dem Feinstaubsensor SDS011. Zusätzlich kann noch ein Sensor für die Luftfeuchtigkeit und Temperatur angeschlossen werden, wie zum Beispiel der DHT22 oder BME280. Geschützt wird die Station von zwei zusammengesteckten Abwasserrohrbogen (Marley Silent HT Bogen). Die Station wird mit einer Firmware des OK

Labs ausgestattet und mit dem WLAN verbunden. Über den Browser wird die Sensorstation eingerichtet. Hierbei wird unteranderem die Frequenz festgelegt, mit der die Station Messungen vornimmt und die Verbindung zu der luftdaten.info-API hergestellt. Über diese Schnittstelle werden die Messdaten mit dem Projekt geteilt. Die Daten sind über eine Karte auf der Website luftdaten.info sichtbar und werden hauptsächlich in CSV-Dateien gespeichert. Hierbei wird eine Datei pro Tag, Station und Sensor erstellt. Alternativ werden die Daten auch im Parquet-Format von Apache angeboten. Apache Parquet ist ein Open-Source, spaltenorientiertes Speicherformat. Diese Dateien enthalten die Daten monatlich von allen Stationen, aber differenzieren noch zwischen den Sensoren. Der grundsätzliche Aufbau der Dateien ist aber identisch.

sensor_id	sensor_type	location	lat	lon	timestamp	P1	durP1	ratioP1	P2	durP2	ratioP2
50	SDS011	31	48.777	9.235	2018-01-01T00:07:06	186.68			177.75		
50	SDS011	31	48.777	9.235	2018-01-01T00:09:34	196.63			187.23		
50	SDS011	31	48.777	9.235	2018-01-01T00:12:03	191.88			182.68		
50	SDS011	31	48.777	9.235	2018-01-01T00:14:32	179.50			170.90		
50	SDS011	31	48.777	9.235	2018-01-01T00:17:06	179.27			170.68		
50	SDS011	31	48.777	9.235	2018-01-01T00:19:41	185.50			176.63		
50	SDS011	31	48.777	9.235	2018-01-01T00:22:11	188.20			179.20		
50	SDS011	31	48.777	9.235	2018-01-01T00:24:39	181.27			172.57		
50	SDS011	31	48.777	9.235	2018-01-01T00:27:08	187.83			178.83		
50	SDS011	31	48.777	9.235	2018-01-01T00:29:37	186.25			177.35		
50	SDS011	31	48.777	9.235	2018-01-01T00:32:05	180.87			172.20		
50	SDS011	31	48.777	9.235	2018-01-01T00:58:59	161.73			153.98		
50	SDS011	31	48.777	9.235	2018-01-01T01:04:20	150.13			142.93		
50	SDS011	31	48.777	9.235	2018-01-01T01:07:00	150.30			143.10		
50	SDS011	31	48.777	9.235	2018-01-01T01:25:30	152.43			145.13		
50	SDS011	31	48.777	9.235	2018-01-01T02:10:40	155.17			147.73		

Abbildung 3: Aufbau der Daten (SDS011)

Wie in Abbildung 3 zu sehen ist, werden zusätzlich zu der Messung und dem Zeitstempel noch Daten zum Ort und zum Sensor hinterlegt. Auch zu sehen sind Spalten ohne Werte. Dies sind Spalten, die für den früheren Feinstaubsensor PPD42NS genutzt wurden. Mittlerweile werden aber fast ausschließlich die SDS011-Sensoren verwendet. Daher sind diese Spalten für diese Arbeit nicht von Relevanz. Auch die Informationen zu dem Sensor sind redundant, da nur der SDS011 in dieser Arbeit betrachtet wird. Als Information für den Ort genügen in dieser Arbeit Längen- und Breitengrade. Allerdings müssen diese nicht zu jeder Messung hinterlegt sein, sondern werden ähnlich einer relationalen Datenbank separat einmal pro Station gespeichert. In dem Zuge wurde pro Station das Start- und Enddatum des Messzeitraums, sowie die Anzahl der Messungen in diesem Zeitraum ermittelt. Für die ersten Visualisierungen werden die Daten auf den Bereich von Deutschland begrenzt. In dieser Arbeit wurde als Zeitraum das Jahr 2018 gewählt. In diesem Jahr war der SDS011 bereits als Feinstaubsensor etabliert und das Projekt hat deutlich an Aufmerksamkeit gewonnen, wie in Abbildung 4a zu sehen ist.

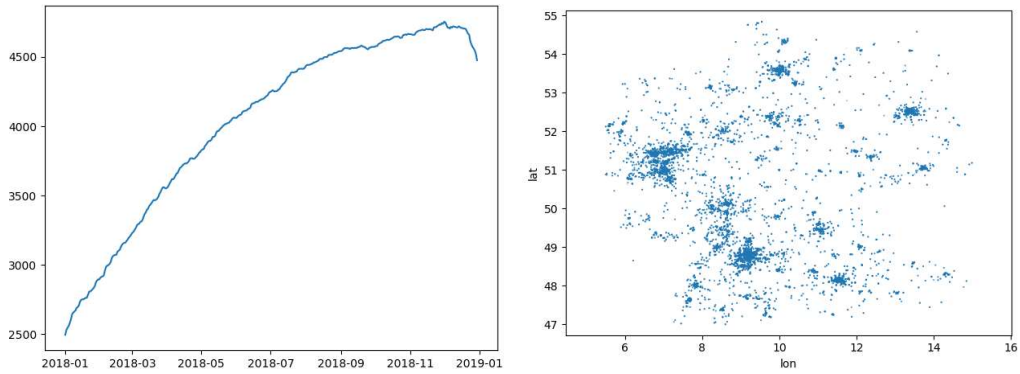

Abbildung 4: SDS011 2018 in Deutschland a) Anzahl der Sensoren, b) Standorte

In Abbildung 4b ist die Verteilung der Sensoren in Deutschland dargestellt. Hierbei ist zu erkennen, dass es Ballungsgebiete gibt. Besonders fällt die Ballung in dem Bereich um Stuttgart auf, da das OK Lab dort am längsten ist und auch vermehrt den mit Sensoren abgedeckten Bereich ausbaut. Auch wenn es weitere Ballungsgebiete gibt, wird in dieser Arbeit der betrachtete Bereich auf Stuttgart eingegrenzt, um einen möglichst aussagekräftigen Bereich für die räumlichen Zusammenhänge zu schaffen.

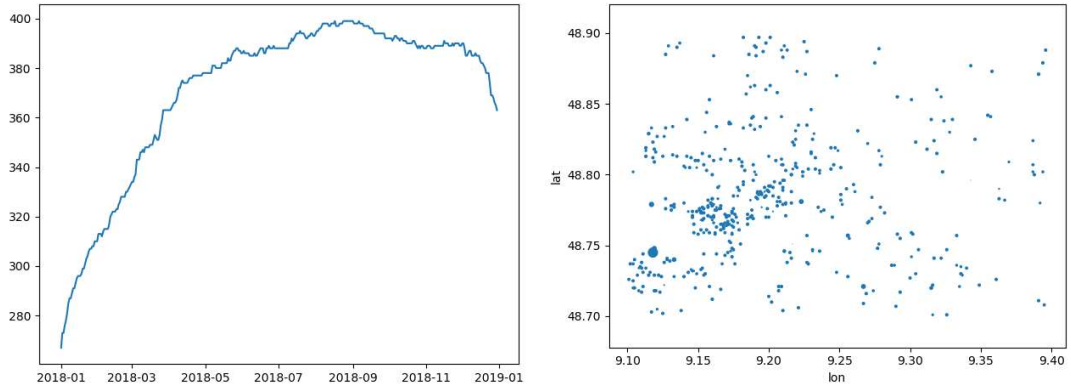

Abbildung 5: SDS011 2018 in Stuttgart a) Anzahl der Sensoren, b) Standorte

In Abbildung 5 sind die beiden Grafiken für Stuttgart dargestellt. Ähnlich zur deutschlandweiten Darstellung ist zu beobachten, dass die Anzahl der Sensoren deutlich zugenommen hat. Allerdings ist auch zu erkennen, dass gegen Ende des Jahres die Anzahl der Sensoren wieder zurück geht. Eine mögliche Erklärung wäre die Deaktivierung von Stationen aufgrund der Wetterlage im Winter. Das konnte aber nicht bestätigt werden. Nur ca. 44,8 % der Sensoren sind über das komplette Jahr aktiv. Dadurch wird nicht nur die zeitliche, sondern auch die räumliche Kontinuität der Daten gestört, da nicht nur an den gleichen Stationen Messungen fehlen, sondern auch Stationen an unterschiedlichen Orten neu installiert oder entfernt werden. Die zeitliche Kontinuität ist auch durch die unterschiedlichen Frequenzen, mit denen die Messungen vorgenommen werden, beeinträchtigt. Damit die DM-Verfahren dennoch konsistente Daten erhalten, muss diese Problematik in der Datenvorverarbeitung behandelt werden.

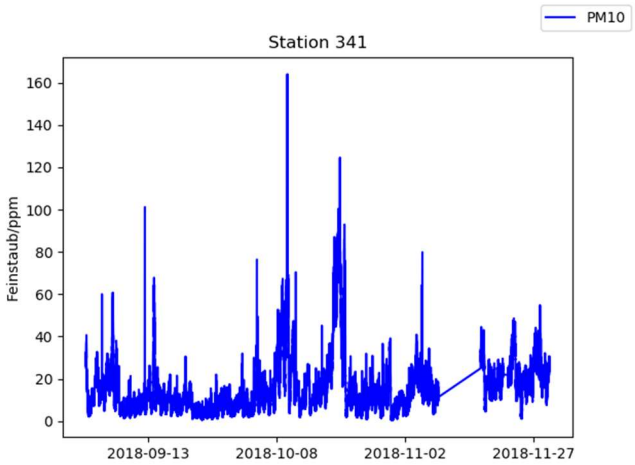

Abbildung 6: Ausschnitt Feinstaubmessungen einer Messstation

In Abbildung 6 ist zu sehen, wie bei Rohdaten üblich, dass die Daten sowohl mit Ausreißern verunreinigt sind als auch Messlücken in den Daten sind. Auch diese Problematik muss in der Datenvorverarbeitung behandelt werden.

3.3 Datenvorverarbeitung

In diesem Abschnitt werden die im vorhergehenden Abschnitt erkannten Problematiken behandelt. In Teilabschnitt 3.3.1 werden als erstes die unterschiedlichen Messfrequenzen vereinheitlicht. Daraufhin werden in Teilabschnitt 3.3.2 sowohl die fehlenden Daten als auch die Ausreißer behandelt.

3.3.1 Datenfrequenz

Grundsätzlich kann für Data-Mining-Verfahren angenommen werden, dass einzelne Datenpunkte unabhängig voneinander sind. Bei Zeitreihenanalysen muss allerdings der zeitliche Zusammenhang beachtet werden. Messungen aus der Vergangenheit haben eine Relevanz und je näher eine Messung an der Gegenwart ist, desto höher ist ihre Bedeutung und Einfluss für Vorhersagen in die Zukunft. Werden mehrere Zeitreihen betrachtet, ist auf die Vergleichbarkeit zu achten. Wie in Abschnitt 3.2 beschrieben, sind die Messreihen der einzelnen Stationen aus zwei Gründen nicht direkt vergleichbar. Einmal beginnen und ggf. enden die Messreihen willkürlich. Wann eine Station an das Netz angeschlossen wird, ist weder vorhersagbar noch definiert. Zweitens ist es jedem Besitzer einer Messstation freigestellt, die Frequenz festzulegen, mit der die Messungen vorgenommen und veröffentlicht werden. Daher kommt es einerseits zu Frequenzen von weniger als einer Messung am Tag, aber auch in anderen Fällen von fast 800 Messungen in der Stunde. Dadurch werden zusätzlich eine große Menge Daten generiert, welche aber kaum einen Mehrwert haben. Die Konzentration von Feinstaub oder anderen Luftdaten ändert sich selten mit einer so hohen Rate. Um eine Vergleichbarkeit zu erreichen, wird eine neue Messfrequenz festgelegt. Diese kann höchstens der Frequenz der Messungen entsprechen:

$$f_{neu} \leq f_{Messung}$$

In diesem Zuge lässt sich auch die Granularität der Daten und damit die Menge der Daten regulieren. In dieser Thesis wurde eine Frequenz von einer Messung pro Stunde gewählt. Hierfür wurden je Station die Messungen über eine Stunde gemittelt und in einem neuen Datensatz abgelegt. In diesem Schritt wurde der Messzeitraum jeder Station auf ein Jahr erweitert und Messzeitpunkte vor und nach der aktiven Zeit wurden mit NaN gekennzeichnet. Durch dieses Vorgehen kann ein Schritt in der Datentransformation für das Kriging eingespart werden.

3.3.2 Datenbereinigung

Wie für Rohdaten üblich, sind diese häufig verunreinigt. Auch in dem Datensatz von luftdaten.info lassen sich einzelne Ausreißer beobachten. Eine Möglichkeit, die Problematik anzugehen, ist die Filterung.
In der Filterung der Daten geht es weniger um die Erkennung dieser Verunreinigung, sondern direkt um die Behandlung. Dies hat zur Folge, dass nicht nur einzelne Werte angepasst werden, viel mehr werden alle Datenpunkte von einer Filterfunktion verändert. Im Folgenden werden zwei gängige Filterfunktionen genauer betrachtet. Beide Funktionen gibt es in einer symmetrischen und asymmetrischen Variante. Hierbei wird unterschieden, ob zum Zeitpunkt der Filterung bereits alle Daten bekannt sind (Offline-Betrieb). In diesem Fall wird die symmetrische Variante genutzt und Daten vor und nach dem betrachteten Datenpunkt werden für die Filterung berücksichtigt. Andernfalls sind bei der asymmetrischen Filterung

die zukünftigen Daten noch nicht bekannt (Online-Betrieb) und folglich können für die Filterung nur vergangene Datenpunkte genutzt werden. Da die genutzten Daten in dieser Arbeit im Nachhinein betrachtet werden, wird im Weiteren auf die symmetrischen Varianten eingegangen.

Symmetrischer gleitender Mittelwert

Der symmetrische gleitende Mittelwert y wird als Fensterfunktion auf die Zeitreihendaten X an der Stelle k angewandt. Die Anzahl der Werte für die Mittelwerts-Berechnung wird durch die Fenstergröße definiert. Die Größe des Fensters ist abhängig von der ungeraden Ordnung $q \in \{3,5,7, \dots\}$.

$$y_k = \frac{1}{q} \sum_{i=k-\frac{q-1}{2}}^{k+\frac{q-1}{2}} x_i$$

Symmetrischer gleitender Median

Die Fensterfunktion des symmetrischen gleitenden Medians ist ähnlich definiert wie bei dem symmetrisch gleitenden Mittelwert. Die Fenstergröße ist wieder abhängig von der ungeraden Ordnung q.

$$m_{kq} \in I_{kq} = \left\{ x_{k-\frac{q-1}{2}}, \dots, x_{k+\frac{q-1}{2}} \right\}$$

mit:

$$\left\| \{ x_i \in I_{kq} | x_i < m_{kq} \} \right\| = \left\| \{ x_i \in I_{kq} | x_i > m_{kq} \} \right\|$$

In den Abbildungen 7 und 8 ist jeweils ein verrauschter Cosinus mit einem Ausreißer im Minimum und die Anwendung beider Filterfunktionen mit unterschiedlicher Fenstergröße. Es ist zu erkennen, dass bei einem größer gewählten Fenster mehr Details rausgefiltert werden. Weiter ist im Vergleich der beiden Filterfunktionen zu sehen, dass der gleitende Mittelwert effektiver das Rauschen filtert, hingegen der gleitende Median den Ausreißer vollständig entfernt. Der Medianfilter wird auch als robuster Filter bezeichnet aufgrund seiner Effektivität gegenüber Ausreißern. Daher wird im weiteren Verlauf der Arbeit dieser Filter für die Bereinigung der Ausreißer verwendet.
Weitere Filter sind der exponentielle Filter, sowie endliche Impulsantwort (*finite impulse response*, FIR) und unendliche Impulsantwort (*infinite impulse response*, IIR) als Tiefpassfilter. Diese Filter lassen sich weiter durch die Filterkoeffizienten dem Aufgabenbereich optimal anpassen (Runkler, 2010 S. 30). In dieser Thesis wird auf diese Filtermöglichkeiten aber nicht weiter eingegangen.

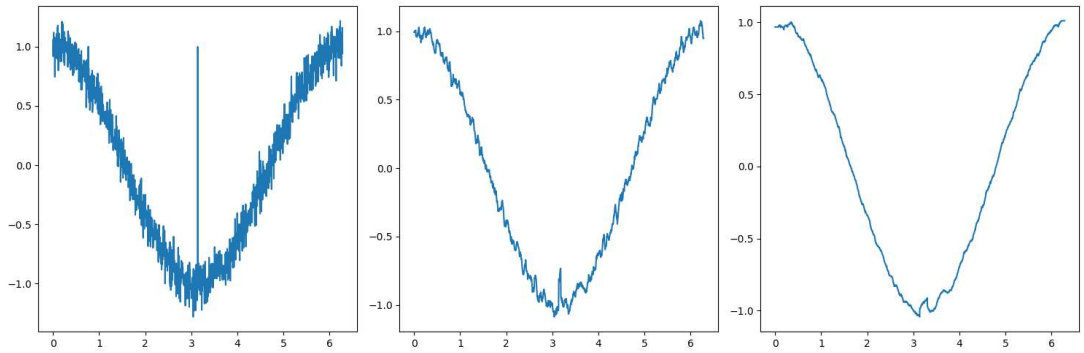

Abbildung 7: Verunreinigter Cosinus, symmetrischer gleitender Mittelwert (q=7), symmetrischer gleitender Mittelwert (q=31)

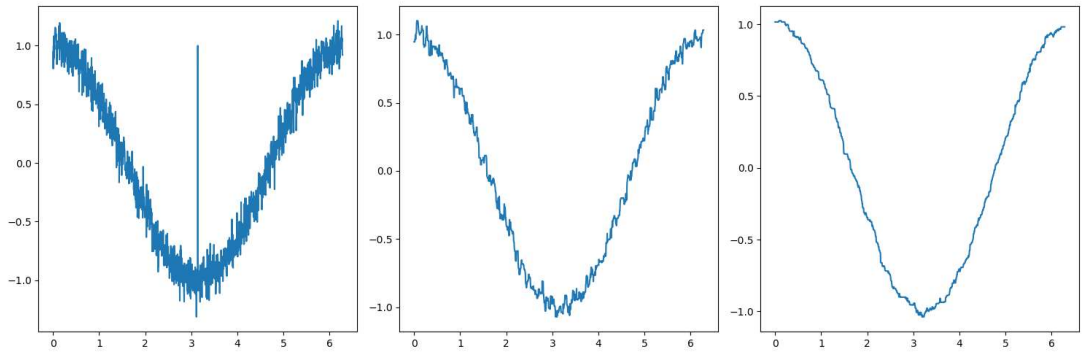

Abbildung 8: Verunreinigter Cosinus, symmetrischer gleitender Median (q=7), symmetrischer gleitender Median (q=31)

Die Filterung bietet eine effektive Möglichkeit, sowohl Rauschen als auch Ausreißer effektiv zu entfernen. Allerdings wird ein komplett neuer Datensatz mit synthetischen Daten erstellt, bei dem ein hoher Detailgrad bereits eliminiert wurde. Um nicht alle Datenpunkte zu verändern, wurde der Datensatz mithilfe der Sigma-Regel auf Ausreißer untersucht. Die Regel wurde ähnlich wie die Filter in einer symmetrischen Fensterfunktion implementiert, um die Ausreißer lokal zu erkennen und die negativen Einflüsse bei schwankenden Daten zu verringern.
Fehlende Daten wurden ebenfalls durch eine symmetrische Fensterfunktion ergänzt. In dieser wurde der Mittelwert der vorhandenen Daten innerhalb des Fensters genutzt. Hierdurch nähert sich der ergänzte Wert den umliegenden Datenpunkten an. Bei großen Intervallen fehlender Daten im Vergleich zur Fenstergröße nähert sich der Wert dem lokalen Mittelwert.

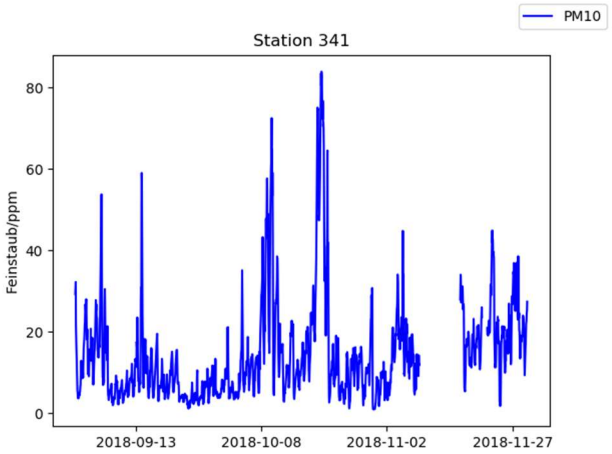

Abbildung 9: stündliche Rohdaten

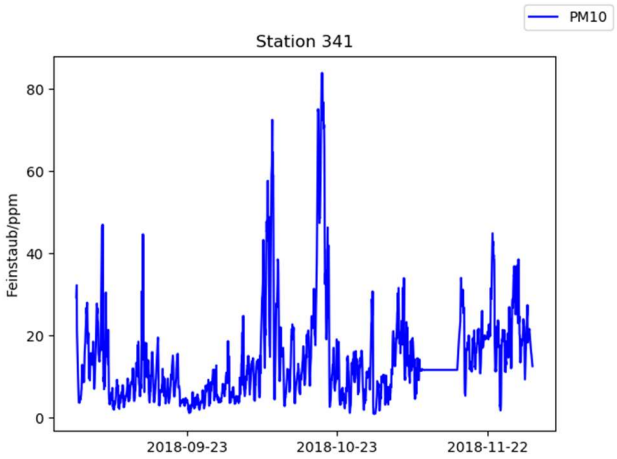

Abbildung 10: ergänzte und bereinigte Daten (q=31, sigma=2)

In Abbildung 9 und 10 sind die Daten vor und nach der Bereinigung von Ausreißern abgebildet. Zum einen ist zu erkennen, dass ein hoher Detailgrad bewahrt wurde. Allerdings wurde die Wirkung starker Ausreißer reduziert, sofern die Änderung der umliegenden Datenpunkte nicht ähnlich ist. Die fehlenden Daten wurden ergänzt. Da ein relativ großes Intervall fehlt, ergibt sich unrealistischerweise ein konstanter Mittelwert. An den Kanten nähert sich der Wert allerdings wieder den umliegenden Messungen. Auch wenn für den Bereich die Daten unrealistisch interpoliert wurden, sind die Werte in einem realistischen Bereich. Zudem ist der Ausschnitt im Vergleich zum gesamten Jahr relativ klein. Daher werden die Daten für den weiteren Verlauf so belassen.
Die fehlenden Daten durch neue und abgeschaltete Stationen werden nicht ergänzt, um keine falschen oder irreführenden Daten zu generieren. Dafür wird akzeptiert, dass weniger Raum-Zeit-Daten für das geostatistische Schätzverfahren verfügbar sind. Dafür bleiben die

Daten fundiert und beruhen auf echten Messungen. Da es vor und nach dem Messzeitraum keine Aufzeichnungen gibt, würde selbst bei einem Regressionsansatz der Schätzfehler mit dem zeitlichen Abstand steigen.

3.4 Zeitliches Data-Mining

Messungen über die Zeit werden als Zeitreihen beschrieben. In dem STDM werden zur Analyse solcher Zeitreihen üblicherweise RNN genutzt. Insbesondere werden LSTM-Netze verwendet, da sie weniger anfällig für *Vanishing* und *Exploding* sind. Im Folgenden wird ein grundsätzliches LSTM-Netz erstellt, um die Nutzbarkeit für die Vorhersage von Feinstaubmessungen an einer Station zu überprüfen.
Für eine verbesserte Lernfähigkeit eines NNs werden die Daten zu Beginn normalisiert. Dadurch liegt der Wertebereich nun zwischen -1…1 und mit einer Standardabweichung von 1. So fallen größere Werte und häufiger vorkommende Werte nicht stärker ins Gewicht.
Im nächsten Schritt wurden die Daten mittels einer Fensterfunktion in Features, die Daten, anhand denen die Vorhersage getroffen werden soll, und das Label geteilt. Hierbei wurden die letzten 24 Messungen als Feature gewählt.
Diese Daten wurden dann in einen Trainingsdatensatz und einen Testdatensatz geteilt in einem Verhältnis von 4:1. Mit dem Testdatensatz kann im Nachhinein überprüft werden, ob ein Overfitting an dem Testdatensatz stattgefunden hat.
Die Wahl der Hyperparameter für ein neuronales Netz kann entscheidende Auswirkungen für den optimalen Erfolg des Trainings haben. Die Wahl der Parameter wurde aus Erfahrung und Ergebnissen aus verwandten Arbeiten getroffen (Proceedings of the 2018 International Conference on Big Data Engineering and Technology, 2018), (Predicting Amazon Spot Prices with LSTM Networks, 2018), (C-LSTM: Enabling Efficient LSTM Using Structured Compression Techniques on FPGAs, 2018). Für die Optimierungsfunktion wurde *adam* (Adam: A Method for Stochastic Optimization, 2014) gewählt. Als Fehlerfunktion wurde die mittlere quadratische Abweichung gewählt (engl. *Mean Squared Error, MSE*). Diese Fehlerfunktion wird häufig für Regressionsprobleme genutzt.
Da in dieser Arbeit kein perfektes Modell entwickelt werden soll, sondern vor allem die Anwendbarkeit auf den Datensatz getestet werden soll, wurden 50 Epochen gewählt, über die das RNN trainiert. Das RNN besteht aus einer LSTM-Schicht gefolgt von einem Dense-Layer (einfache Neuronen Schicht) mit einem Neuron. Die Aktivierungsfunktion von der LSTM-Schicht ist *tanh*. Das ist die Standardeinstellung in Keras (Ker19).

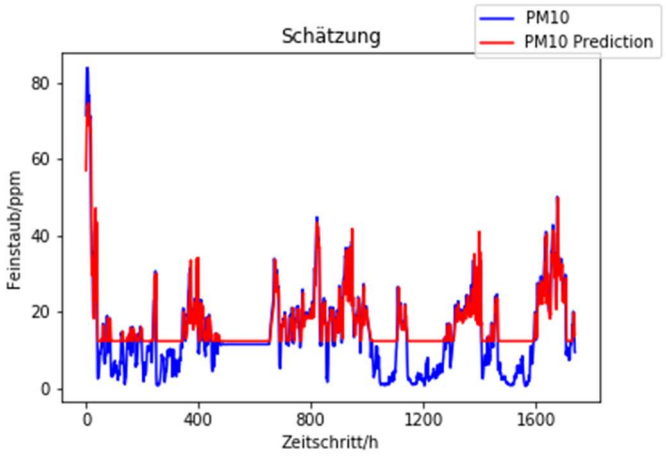

Abbildung 11: Vorhersagen für PM10 mit ReLu als Aktivierungsfunktionen

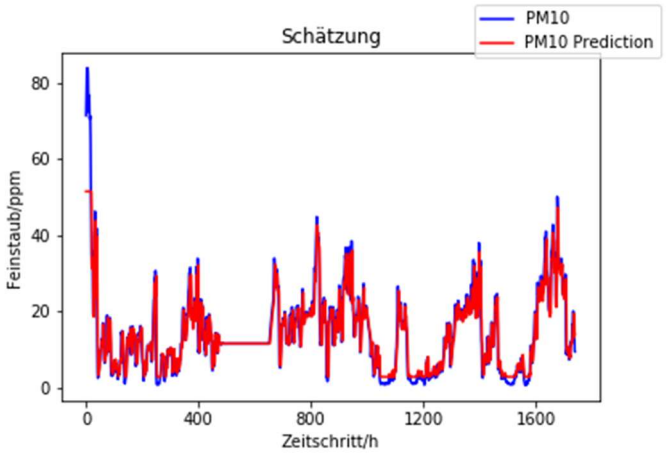

Abbildung 12: Vorhersagen für PM10 mit linearer Aktivierungsfunktionen

In Abbildung 11 und 12 sind die Schätzungen des Modells anhand der Testdaten abgebildet. Zuerst wurde als Aktivierungsfunktion für den Dense-Layer *ReLu* (*Rectified Linear Unit*) gewählt, da diese Funktion für NN gängig ist. Allerdings wurden alle Daten unterhalb eines gewissen Grenzwerts weggeschnitten, was durch die Funktionsweise von *ReLu* zu erklären ist. Daraufhin wurde noch eine lineare Aktivierungsfunktion gewählt. Diese zeigt für den Netzaufbau die besseren Ergebnisse. Allerdings wird der hohe Ausschlag zu Beginn nicht so gut vorhergesagt.

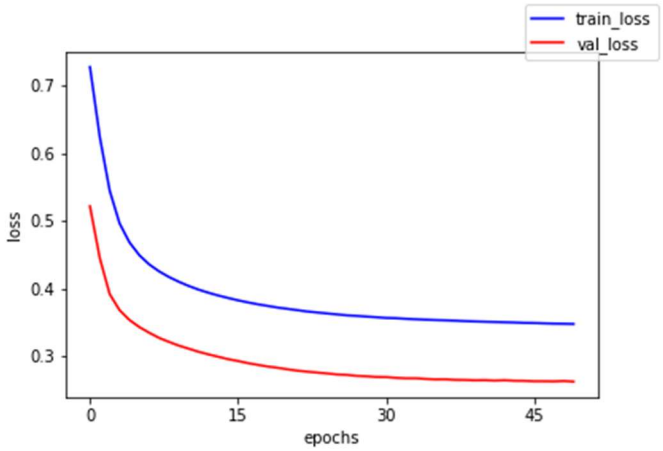

Abbildung 13: Verlustfunktion (ReLu)

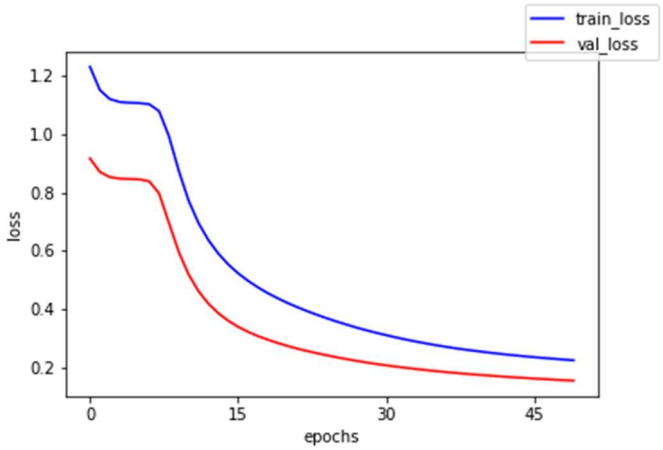

Abbildung 14: Verlustfunktion (linear)

In Abbildung 13 und 14 ist der Verlauf der Verlustfunktion über die Epochen abgebildet. Hier ist zu erkennen, dass die lineare Aktivierungsfunktion insgesamt besser abschneidet. Auch nach drei Wiederholungen zeigten sich ähnliche Ergebnisse. Weiter ist zu erkennen, dass es noch nicht zu einem Overfitting gekommen ist, da der Verlauf der Verlustfunktion an den Testdaten (val_loss) stets unterhalb der an den Trainingsdaten (train_loss) ist. Allerdings ist das hier angewendete Modell noch sehr einfach. Bei einem komplexeren Modell ist die Gefahr von Overfitting höher (Tetko, et al., 1995 S. 826).

3.5 Spatial Data-Mining

Die Messorte in dem betrachteten Datensatz sind sehr unregelmäßig verteilt. Mittels Kriging können Werte an Orten geschätzt werden, die zwischen den Messpunkten liegen. Dies bietet die Möglichkeit, Schätzungen für ein regelmäßiges Raster vorzunehmen.

Hierfür wurden zunächst die Daten transformiert. Die Daten, die derzeit stündlich pro Station vorliegen, werden nun pro Zeitschritt gesammelt und mit den Koordinaten der jeweiligen Station versehen. Für das weitere Vorgehen wird ein einzelner Zeitschritt am Ende des Jahres betrachtet, da dort eine große Anzahl an Messstationen aktiv war.

3.5.1 Variografie

Als erstes werden die Schritte der Variografie durchlaufen. Hierbei wird der räumliche Zusammenhang nach der Theorie der regionalisierten Variablen durch eine Autokorrelationsstruktur γ dargestellt. Diese Struktur muss, wie nachfolgend gezeigt, geschätzt werden:

$$\gamma^*_{(h)} = \frac{1}{2} \cdot \frac{\sum_{i=1}^{n(h)} \big(z(x) - z(x+h)\big)^2}{n(h)}$$

Hierbei wird die Varianz der Zufallswerte z an den Punkten x innerhalb einer festgelegten Entfernung h (*Lag*) bestimmt und durch die Anzahl dieser Wertepaare n dividiert.
Das ermittelte empirische Variogramm muss daraufhin durch ein mathematisches Modell beschrieben werden. Die Wahl der Modelle ist eingeschränkt, da sie bestimmte Eigenschaften erfüllen müssen, damit die Lösbarkeit der Kriging-Gleichung gegeben ist (siehe Teilabschnitt 2.4.1).
Folgende Modelle sind die gebräuchlichsten:

Sphärisches Modell:

$$\gamma(h) = C_0 + C\left(\frac{3h}{2a} - \frac{h^3}{2a^3}\right) \qquad \forall h \leq a \text{ und}$$

$$\gamma(h) = C_0 + C \qquad \forall h > a$$

Exponentielles Modell:

$$\gamma(h) = C_0 + C\left(1 - \exp\left[\frac{-h}{a}\right]\right)$$

Gaussches Modell:

$$\gamma(h) = C_0 + C\left(1 - \exp\left[\frac{-h^2}{a^2}\right]\right)$$

Der Verlauf der Modelle ist in Abbildung 15 abgebildet.

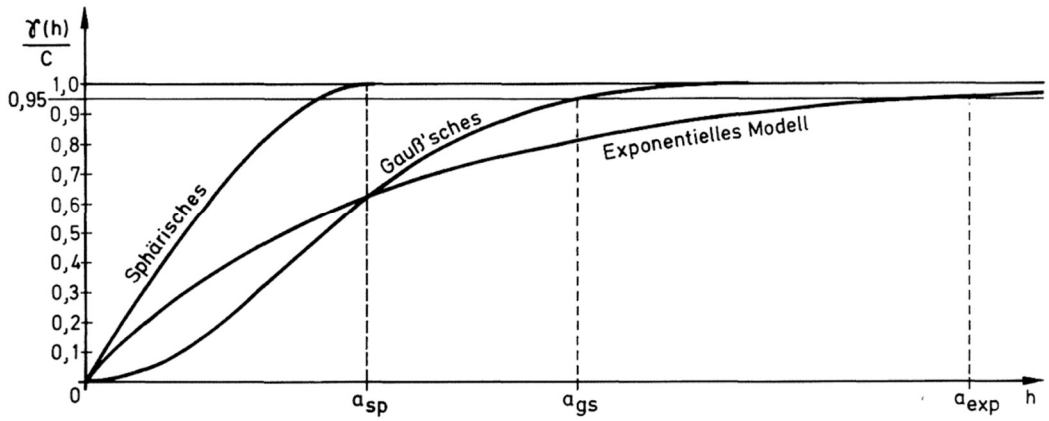

Abbildung 15: Variogramm Modelle (Akin et al. 1988 S. 45 Abb. 2.3.5)

Für das experimentelle Variogramm wurden nun für einen Zeitpunkt die Varianz der Messwerte und die Entfernung zueinander bestimmt. Da die Anzahl der Stationen in dem betrachteten Zeitraum stark zunimmt, wurde ein Zeitpunkt gewählt, in dem möglichst viele Stationen aktiv waren. In dem betrachteten Bereich um Stuttgart sind zu dem betrachteten Zeitpunkt 355 Messstationen aktiv. Nun wurden die ermittelten Werte in „Lags“ aufgeteilt. Lags beschreiben dabei einen Entfernungsbereich zwischen zwei Stationen. In dieser Arbeit wurde der maximal betrachtete Entfernungsbereich auf die Hälfte der Diagonalen des betrachteten Bereichs festgelegt (zitiert nach JOURNEL, A. G. & HUUBREGTS, C. J. (1978): Mining Geostatistics - New York, vgl. (Schroder, et al., 1998 S. 152)). Die Entfernung während der Berechnungen wurde anhand der Koordinatenpunkte berechnet. Bei einer diagonalen Strecke zwischen den beiden Eckkoordinaten von ca. 26,5 km werden in dem experimentellen Variogramm nur Messungen mit einem maximalen Abstand von etwa 13,25 km betrachtet. Dieser Bereich wurde in 200 Lags unterteilt. Damit beschreibt ein Lag ein Entfernungsfenster über ungefähr 66 m. Der erste Lag beschreibt dadurch eine Entfernung von 0 m-66 m, der zweite von 66 m-132 m und nach dem gleichen Prinzip weiter bis zu dem letzten Lag von 13,184 km-13,25 km. Den Lags wurden nun die Varianzen zu geordnet und die Variogrammwerte $\gamma^*_{(h)}$ bestimmt.

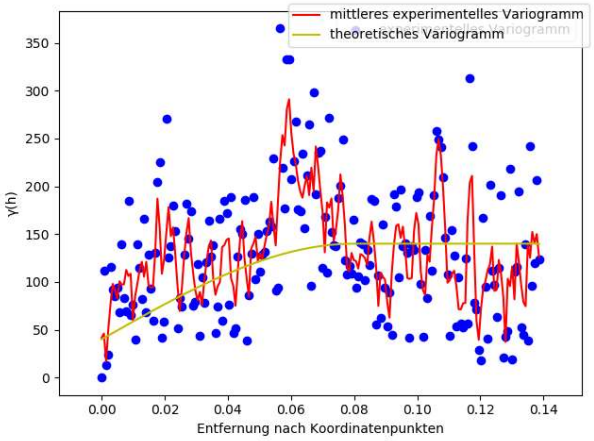

Abbildung 16: Variogramm

In Abbildung 16 ist das resultierende Variogramm abgebildet. Zusätzlich wurde der laufende Mittelwert eingetragen, um die Änderung des Variogramms über die Lags zu verdeutlichen. Auch wenn große Schwankungen zu erkennen sind, sind zu Beginn noch geringere Variogrammwerte zu erkennen, während mit zunehmender Entfernung die Schwankung steigt. Dazu wurde ein möglichst passendes Variogramm-Modell gewählt. Mit einem Sill C von etwa 140 bei einer Range a von 0.08 und einem Nuggeteffect $C(0)$ von etwa 40. Diese experimentell ermittelten Parameter werden nun in das theoretische Variogramm übertragen. Da die Form nicht direkt auf ein Modell schließen lässt, wurde das sphärische Modell gewählt. Mit dieser Beschreibung der Autokorrelation kann nun in dem nächsten Schritt, dem Kriging, Schätzungen vorgenommen werden.

3.5.2 Kriging

Mithilfe des Kriging können entweder bestimmte Lokalitäten gewählt werden, die eventuell von besonderem Interesse sind oder es wird ein gleichmäßiges Schätzraster über einem Bereich definiert, um eine allgemeinere Betrachtung zu haben und eine räumliche Kontinuität zu schaffen. Die räumliche Kontinuität bildet in dieser Arbeit den größten Vorteil. Sie ermöglicht es, trotz einer schwankenden Menge an Messorten eine konstante Menge an Datenpunkten zu schaffen.
Ebenfalls, wie bei den zeitlichen Messintervallen, lässt sich über die Größe und Granularität des Rasters die Performanz der genutzten DM-Verfahren beeinflussen. Es ist zwischen dem Nutzen für die Ziele der Aufgabenstellung und der geforderten Leistung an das Modell abzuwägen, da die Schätzungen vieler Rasterpunkte zu hohem Rechenaufwand führen.

In dieser Arbeit wird ein Raster mit hundert mal hundert Schätzpunkten genutzt. Dieses wird über dem Bereich Stuttgart aufgespannt, um eine möglichst gute Schätzung vornehmen zu können.

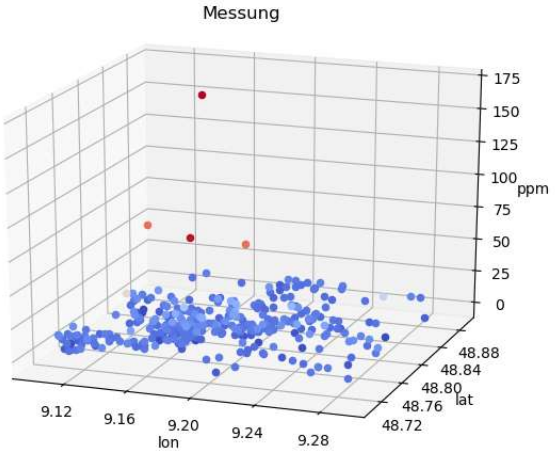

Abbildung 17: Messungen in Stuttgart

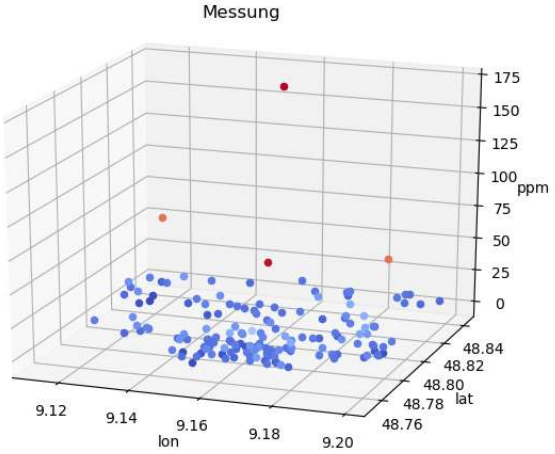

Abbildung 18: Messungen in Stuttgart (reduziert)

In Abbildung 17 ist der komplette betrachtete Bereich in Stuttgart zu sehen und in Abbildung 18 ist der Ausschnitt gezeigt, in dem das zu schätzende Raster gelegt ist. Der Bereich ist ca. 10 km lang und 7 km breit. Dementsprechend sind die Schätzpunkte in etwa 100 m - 70 m voneinander entfernt. Die Schätzung wird durch das Ordinary-Kriging vorgenommen und wird in den folgenden Teilabschnitten beschrieben.

Grundsätzlich lässt sich die Form des Schätzers folgendermaßen beschreiben:

$$z^*(x) = \sum_{i=1}^{n} \lambda_i \, z(x_i)$$

Aufgrund der Unverzerrtheitsbedingung ist die Summe der Gewichte λ über alle Messpunkte n dabei eins.
Wegen der bereits genannten Eigenschaften des Kriging-Schätzers und des Variogramms ergibt sich für die Schätzvarianz σ^2, wobei x_0 der zu schätzende Datenpunkt ist:

$$\sigma^2(x_0) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i, x_j)\right) - 2 \sum_{i=1}^{n} \lambda_i \, \gamma(x_i, x_0)$$

Die Varianz soll minimiert werden. Hierfür werden nach dem Lagrange-Prinzip die partiellen Ableitungen nach den Gewichten gebildet und gleich null gesetzt. Die Einführung des Lagrange-Parameters μ ermöglicht das Lösen des entstandenen Gleichungssystems (Gau, 2010 S. 35). Dieses hat durch die Unverzerrtheitsbedingung $n + 1$ Gleichungen und wird als Kriging-System bezeichnet:

$$\sum_{j=1}^{n} \lambda_i \gamma(x_i, x_j) + \mu = \gamma(x_i, x_0)$$

Vereinfacht dargestellt in Matrixschreibweise:

$$A\lambda = b$$

Wobei die Matrix A die Werte aus dem theoretischen Variogramm enthält. Der Spaltenvektor λ enthält die geschätzten Gewichte und den Lagrange-Parameter μ. Der Spaltenvektor b enthält die geschätzten Variogrammwerte zwischen den bekannten Datenpunkten und dem zu schätzenden Datenpunkt:

$$A = \begin{bmatrix} \gamma(x_1, x_1) & \dots & \gamma(x_n, x_1) & 1 \\ \vdots & \ddots & \dots & \vdots \\ \gamma(x_1, x_n) & \dots & \gamma(x_n, x_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}, \qquad \lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} \; und \; b = \begin{bmatrix} \gamma(x_1, x_0) \\ \vdots \\ \gamma(x_n, x_0) \\ 1 \end{bmatrix}.$$

Daher lassen sich wie folgt die Gewichte berechnen:

$$\lambda = A^{-1} b$$

Die minimale Schätzvarianz, die auch Kriging-Varianz genannt wird, lässt sich wie folgt bestimmen:

$$\sigma_k^2(x_0) = \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_0) + \mu = (x_i, x_0)$$

Und in Matrixschreibweise:

$$\sigma_k^2(x_0) = b^T \lambda = b^T A^{-1} b$$

Als erstes wurde die Kriging-Gleichung in Matrixform aufgestellt. Für die Matrix A wurde mittels des theoretischen Variogramms der Variogrammwert zwischen jedem Messpunkt untereinander bestimmt. Dies hat die Aufgabe, Redundanzen zwischen Messpunkten hinsichtlich des Schätzpunktes zu minimieren. Als nächstes wurde der Spaltenvektor b bestimmt. Hierfür wurde der Variogrammwert zwischen jedem Messpunkt und dem Schätzpunkt berechnet. Durch den Spaltenvektor wird die Nähe zu dem Schätzpunkt gewichtet. Nun wurden die Matrix und der Spaltenvektor nach dem Lagrange-Prinzip aufbereitet. Durch das Lösen des Gleichungssystems, in dem die Inverse der Matrix mit dem Spaltenvektor multipliziert wird, erhält man die Gewichte der Messpunkte in dem Spaltenvektor λ. Dieser wurde daraufhin mit den Messungen der Stationen multipliziert. Zusätzlich wurde die Kriging-Varianz σ_K^2 bestimmt, indem der transponierte Spaltenvektor b mit dem Spaltenvektor λ multipliziert wurde.
Für den betrachteten Zeitpunkt ergeben sich daraufhin folgende Abbildungen:

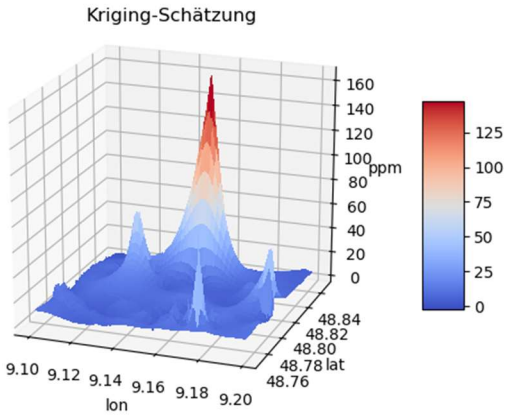

Abbildung 19: Kriging-Schätzung

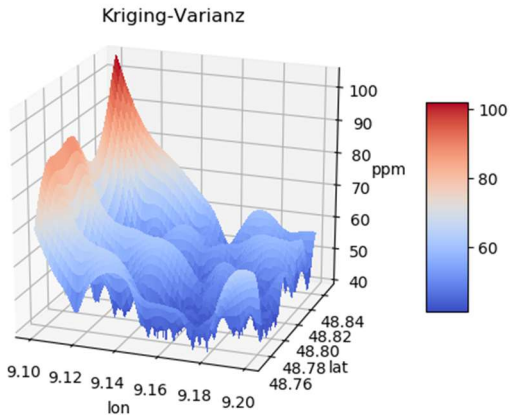

Abbildung 20: Kriging-Varianz

In Abbildung 19 sind die Kriging-Schätzungen abgebildet. Im Vergleich mit den vorhandenen Messungen lassen sich klare Bezüge herstellen. Einerseits ist ein Großteil der Schätzungen in einem Bereich von 0-25 ppm, aber auch die vier höheren Werte sind abgebildet. In Abbildung 20 ist die Kriging-Varianz dargestellt. Auch hier lassen sich Bezüge zu den vorhandenen Messungen herstellen. Das Schätzraster grenzt bei dem Längengrad von 9,10 ° an den Rand der verfügbaren Messungen. Daher lassen sich die Schätzungen nur auf einer geringeren Anzahl an Messwerten stützen. Auch wenn die Schätz-Varianz relativ hoch im Vergleich zu den Schätzungen ist, zeigen die Schätzungen ein Bild, welches sehr nah an der Erwartung ist.

3.6 Fazit

Ziel dieser Arbeit war es einerseits den Datensatz von luftdaten.info zu untersuchen und für die Nutzbarkeit für Data-Mining Verfahren sowohl im zeitlichen als auch räumlichen Kontext zu beurteilen und daraufhin einige raumzeitliche Data-Mining-Verfahren anzuwenden.
Der Datensatz zeichnet sich durch eine hohe Variabilität aus. Das begründet sich vor allem an dem wachsenden Interesse an dem Projekt und der daraus resultierenden wachsenden Anzahl an Messstationen. Auch die frei wählbare Frequenz, mit der die Messungen vorgenommen werden, erschwert die Vergleichbarkeit. Diese wurde durch die Definition einer globalen Datenfrequenz auf einen Datenpunkt pro Stunde wiederhergestellt. Hierbei wurde aber auch ein hoher Detailgrad aufgegeben. Die Daten enthalten sowohl Ausreißer als auch gelegentliche Aussetzer der Messungen. Die Ausreißer wurden mittels der Sigma-Regel in einer Fensterfunktion lokal erkannt und durch den lokalen Median ersetzt. Dadurch wurden auffällige Ausreißer eliminiert, ohne den gesamten Datensatz zu verändern. Das Auffüllen der fehlenden Werte mit dem lokalen Mittelwert hat den Vorteil, einen möglichst

passenden Wert im Kontext der anderen zu geben. Allerdings ergeben sich bei längeren Ausfällen unrealistische Werte. Hierbei könnten Schätzungen durch Kriging eine Möglichkeit bieten, die Daten durch das Umfeld zu rekonstruieren.
Nachdem die Daten vorverarbeitet waren, wurden sie für das Long short-term memory-Netz transformiert. Die Ergebnisse zeigen, dass sich relativ gute Prognosen mit den Daten und einem vergleichsweise einfachen rekurrenten neuronalen Netz erzielen lassen. Allerdings können die Ergebnisse vermutlich noch durch eine präzisere Wahl der Hyperparameter und einen komplexeren Aufbau des Netzes verbessert werden. In der genutzten Konfiguration kam es nicht zum Overfitting, vermutlich durch die Einfachheit des Modells.
Die Anwendung des geostatistischen Schätzverfahrens mit dem Kriging-Algorithmus hat gute Ergebnisse erbracht, obwohl das theoretische Variogramm grob gewählt und die Entfernungsberechnung über die Geokoordinaten vereinfacht wurde. Bei dieser Anwendung hat die vergleichsweise hohe Dichte an Messstationen zu den guten Ergebnissen stark beigetragen. In anderen Teilen Deutschlands mit geringerer Abdeckung von Messstationen muss vermutlich ein deutlich präziseres Variogramm-Modell gewählt werden. In diesem Fall sollten die Kriging-Schätzungen mittels der k-fachen-Kreuzvalidierung überprüft werden. Bei großräumigeren Analysen sollte auch die Höhe der Messstation berücksichtigt werden, da sowohl die Entfernung dadurch beeinflusst wird, aber auch andere Gegebenheiten in der Luft herrschen.

4 Ausblick

Durch die Analyse und Experimente wurden viele wertvolle Erkenntnisse gewonnen und dienen als Grundlage für weiterführende Experimente und komplexere raumzeitliche Data-Mining-Anwendungen.
Allerdings bewegte sich die Arbeit aufgrund von technischen Möglichkeiten nur in einem begrenzten Bereich. Durch Berechnung auf leistungsstarken Servern könnten zum Beispiel beim Kriging größere als auch feinere Rasterungen berechnet werden. Ebenso könnten die Möglichkeiten der Long short-term memory-Netze noch besser ausgereizt werden, da diese es ermöglichen, bis zu 1000 Zeitschritte effektiv zu bearbeiten (Hochreiter, et al., 1997 S. 1736).
Wie bereits in Teilabschnitt 3.6 erwähnt, sollte für weitere Analysen vor allem in weniger gut bemessenen Gebieten das Kriging Modell noch genauer validiert werden.
In dieser Arbeit wurden ausschließlich die Feinstaubmessungen begutachtet. Allerdings könnte durch das Hinzufügen von zusätzlichen Datenkanälen mit anderen atmosphärischen Messungen ggf. bessere oder neue Erkenntnisse gewonnen werden. Insbesondere Messungen zur Luftfeuchtigkeit, Luftdruck, Temperatur können einen Mehrwert bieten. Das Landesamt für Umwelt, Messungen und Naturschutz Baden-Württemberg (LUBW) hat den SDS011 auf seine Eignung geprüft. Hierbei stellte sich heraus, dass der Sensor nur bei einer mittleren Luftfeuchtigkeit (50 – 70 %) zufriedenstellende Ergebnisse liefert. Weiter kommt es zu deutlichen Abweichungen bei Schwankungen von Luftfeuchte, Temperatur und Luftdruck (LUBW, 2017 S. 20). Diese Daten sind teilweise auch mit in dem Datensatz vorhanden, können andernfalls aber auch über den Deutschen Wetterdienst (DWD) bezogen werden.
Um zusätzliche Einflüsse wie Wind und Höhe in das geostatistische Schätzverfahren einzubringen, kann *Kriging with external Drift* genutzt werden (Hudson, et al., 1994). Diese Erweiterung kann einen starken Einfluss auf die Ergebnisse haben, da sich zum Beispiel innerhalb eines Talkessels ein eigenes Mikroklima bilden kann, während auf dem flachen Land durch Winde ein homogeneres Klima herrscht.
Durch die Vereinheitlichung des Messrasters und der Zeitschritte können Spatiotemporal-Raster generiert werden. Hierbei ist zu überprüfen, ob das Variogramm für jeden Zeitschritt passend ist und muss ggf. angepasst werden. Spatiotemporal-Raster bilden einen weiteren Datentyp im raumzeitlichen Data-Mining und ermöglichen die Nutzung vieler weiterer Verfahren. Mit Clusteranalysen in diesem Datentyp können Datenähnlichkeiten in Raum und Zeit ermittelt werden (Liu, et al., 2013). Hierdurch ließe sich auch die Bewegung zum Beispiel von Feinstaubwolken verfolgen. Durch die Messung der Bewegung ließen sich die Daten auch noch in Datenverläufe (engl. *Trajectories*), den letzten typischen Datentyp in raumzeitlichen

Data-Mining, umwandeln. Diese beschreiben die Wege von Objekten und Messungen in Raum und Zeit, wie zum Beispiel Satellitenlaufbahnen oder Bewegungen von Bojen.
Gleichmäßige Raster eignen sich auch für die Analyse mit *Convolutional neuronal Networks* (CNN) (Large-scale Video Classification with Convolutional Neural Networks, 2014). Diese werden häufig in der Bildverarbeitung verwendet, wobei sie die örtlichen Zusammenhänge der Bilddaten verarbeiten. Auf diese Weise können auch Messdaten im Raum wie ein Bild analysiert werden.

Literaturverzeichnis

Adam: A Method for Stochastic Optimization. Kingnma, Diederik P. und Ba, Jimmy. 2014. 2014.

Bellec, Pierre, et al. 2006. Identification of large-scale networks in the brain using fMRI. NeuroImage. 2006, Bd. 29, 4, S. 1231 - 1243.

Bhattacharjee, S., Mitra, P. und Ghosh, S. K. 2014. Spatial Interpolation to Predict Missing Attributes in GIS Using Semantic Kriging. IEEE Transactions on Geoscience and Remote Sensing. 2014, Bd. 52, 8, S. 4771-4780.

Blumensath, Thomas, et al. 2013. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. NeuroImage. 2013, Bd. 76, 1, S. 313 - 324.

Brunsdon, Chris, Fotheringham, A. Stewart und Charlton, Martin E. 1996. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis. 1996, Bd. 28, 4, S. 281-298.

Celik, M., et al. 2008. Mixed-Drove Spatiotemporal Co-Occurrence Pattern Mining. IEEE Transactions on Knowledge and Data Engineering. 2008, Bd. 20, 10, S. 1322 - 1335.

Chollet, François. 2018. Deep Learning mit Python und Keras. [Übers.] aus dem Amerikanischen von Knut Lorenzen. Frechen : mitpVerlag GmbH & Co. KG, 2018.

Cleve, Jürgen und Lämmel, Uwe. 2016. Data Mining. Berlin/Boston : Walter de Gruyter GmbH, 2016.

C-LSTM: Enabling Efficient LSTM Using Structured Compression Techniques on FPGAs. Wang, Shuo, et al. 2018. [Hrsg.] Association for Computing Machinery. New York, NY, USA : s.n., 2018. S. 11–20.

Craddock, R. Cameron, et al. 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human Brain Mapping. 2012, Bd. 33, 8, S. 1914-1928.

Discovery of Climate Indices Using Clustering. Steinbach, Michael, et al. 2003. New York, NY, USA : Association for Computing Machinery, 2003. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. S. 446–455. 1581137370.

Ester, M., et al. 1996. A density-based algorithm for discovering clusters. KDD. 1996, Bd. 96, S. 226–231.

Ester, Martin und Sander, Jörg. 2000. Kowledge Discovery in Databases Techniken und Anwendungen. Berlin Heidelberg : Springer-Verlag, 2000.

Fayyad, Usama, Piatetsky-Shapiro, Gregory und Smyth, Padhraic. 1996. From Data Mining to Knowledge Discovery in Databases. AI Magazine. 17, 1996, 3, S. 37-54.

Gau, Christian. 2010. Geostatistik in der Baugrundmodellierung. Dissertation. 2010.

Géron, Aurélien. 2018. Praxiseinstieg Machine Learning mit Scikit-Learn und Tensorflow. [Übers.] Aus dem Englischen von Kristian Rother. Heidelberg : dpunkt.verlag GmbH, 2018.

Giudici, Paolo. 2010. Data Mining Model Comparison. [Buchverf.] Maimon Oded und Lior Rokach. Data Mining and Knowledge Discovery Handbook. 2. New York Dordrecht Heidelberg London : Springer, 2010, S. 641-654.

Goutte, Cyril, et al. 1999. On Clustering fMRI Time Series. NeuroImage. 1999, Bd. 9, 3.

Grzymala-Busse, Jerzy W. und Grzymala-Busse, Witold J. 2010. Handling Missing Attribute Values. [Buchverf.] Oded Maimon und Lior Rokach. Data Mining and Knowledge. 2. New York Dordrecht Heidelberg London : Springer, 2010, S. 22-52.

Heinrich, Uwe. 1994. Flächenschätzung mit geostatistischen Verfahren - Variogrammanalyse und Kriging. [Buchverf.] Winfried Schröder, Lutz Vetter und Otto (Hrsg.) Fränzle. Neuere statistische Verfahren und Modelbildung in der Geoökologie. Braunschweig/Wiesbaden : Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, 1994, S. 144-165.

Heller, Ruth, et al. 2006. Cluster-based analysis of FMRI data. NeuroImage. 2006, Bd. 33, 2, S. 599 - 608.

Hochreiter, Sepp und Schmidhuber, Jürgen. 1997. Long Short-Term Memory. Neural Computation. 1997, Bd. 9, 8, S. 1735-1780.

Huang, Kaizhu, et al. 2019. Deep Learning: Fundamentals, Theory and Application. Schweiz : Springer Nature Switzerland AG, 2019.

Hudson, Gordon und Wackernagel, Hans. 1994. Mapping temperature using kriging with external drift: Theory and an example from scotland. International Journal of Climatology. 1994, Bd. 14, 1, S. 77-91.

Jarvis, R. A. und Patrick, E. A. 1973. Clustering Using a Similarity Measure Based on Shared Near Neighbors. IEEE Transactions on Computers. 1973, Bde. c-22, 11, S. 1025-1034.

Jiang, Zhe und Shekhar, Shashi. 2017. Spatial Big Data Science. Schweiz : Springer International Publishing AG, 2017.

Karypis, G., Han, Eui-Hong und Kumar, V. 1999. Chameleon: hierarchical clustering using dynamic modeling. Computer. 1999, Bd. 32, 8, S. 68-75.

Kelejian, Harry H. und Prucha, Ingmar R. 1999. A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. International Economic Review. 1999, Bd. 40, 2, S. 509-533.

Keras Documentation. [Online] Community-Projekt, initiiert durch François Chollet.[Zitat vom: 9. August 2019.] https://keras.io/layers/recurrent/.

Large-scale Video Classification with Convolutional Neural Networks. Karpathy, Andrej, et al. 2014. [Hrsg.] IEEE. 2014. 2014 IEEE Conference on Computer Vision and Pattern Recognition. S. 1725-1732.

Liu, Xiao, Chang, Catie und Duyn, Jeff. 2013. Decomposition of Spontaneous Brain Activity into Distinct fMRI Co-activation Patterns. Frontiers in Systems Neuroscience. 2013, Bd. 7, S. 101.

Liu, Xiao, Chang, Catie und H. Duyn, Jeff. 2018. Decomposition of spontaneous brain activity into distinct fMRI co-activation patterns. [Buchverf.] Emili Balaguer-Ballester, et al. Metastable Dynamics of Neural Ensembles. s.l. : Frontiers Media SA, 2018.

Lu, Yingli, Jiang, Tianzi und Zang, Yufeng. 2003. Region growing method for the analysis of functional MRI data. NeuroImage. 2003, Bd. 20, 1, S. 455 - 465.

LUBW. 2017. Messungen mit dem Feinstaubsensor SDS011 Ein Vergleich mit einem eignungsgeprüften Feinstaubanalysator. [Online] 2017. [Zitat vom: 04. 08 2019.] http://www4.lubw.baden-wuerttemberg.de/servlet/is/268831/messungen_mit_dem_feinstaubsensor_sds011.pdf?command=downloadContent&filename=messungen_mit_dem_feinstaubsensor_sds011.pdf.

Manaswi, Navin Kumar. 2018. Deep Learning with Applications Using Python. Bangalore, Karnataka, India : Apress, 2018.

Mezer, Aviv, et al. 2009. Cluster analysis of resting-state fMRI time series. NeuroImage. 2009, Bd. 45, 4.

Oliver, M. A. und Webster, R. 1990. Kriging: a method of interpolation for geographical information systems. [Hrsg.] Taylor & Francis. International Journal of Geographical Information Systems. 1990, Bd. 4, 3, S. 313-332.

Oliver, Margaret A. und Webster, Oliver. 2015. Basic Steps in Geostatistics: The Variogramm an Kriging. Cham Heidelberg New York Dordrecht London : Springer, 2015.

Predicting Amazon Spot Prices with LSTM Networks. Baughman, Matt, et al. 2018. [Hrsg.] Association for Computing Machinery. New York, NY, USA : s.n., 2018. Proceedings of the 9th Workshop on Scientific Cloud Computing. S. 1–7.

Proceedings of the 2018 International Conference on Big Data Engineering and Technology. Zhang, Zhenkun, et al. 2018. [Hrsg.] Association for Computing Machinery. New York, NY, USA : s.n., 2018. Proceedings of the 2018 International Conference on Big Data Engineering and Technology. S. 73–77.

Runkler, Thomas A. 2010. Data Mining Methoden und Algorithmen intelligenter Datenanalyse. Wiesbaden : Vieweg+Teubner | GWV Fachverlage GmbH, 2010.

Sartorius, Gerhardt. 2019. Erfassen, Verarbeiten und Zuordnen multivariater Messgrößen: Neue Rahmenbedingungen für das Nächste-Nachbarn-Verfahren. Hagen : Springer-Verlag, 2019.

Schroder, M., et al. 1998. Spatial information retrieval from remote-sensing images. II. Gibbs-Markov random fields. IEEE Transactions on Geoscience and Remote Sensing. 1998, Bd. 36, 5, S. 1446-1455.

Shepherd, Mike. 2009. Oil Field Production Geology. Tulsa, Oklahoma : The American Association of Petroleum Geologists, 2009.

Spatio-Temporal Data Mining: A Survey of Problems and Methods. Atluri, Gowtham, Karpatne, Anuj und Kumar, Vipin. 2018. 51, New York, NY, USA : Association for Computing Machinery, 2018, Bd. ACM Comput. Surv. 4.

Tetko, Igor V., Livingstone, David J. und Luik, Alexander I. 1995. Neural network studies. 1. Comparison of overfitting and overtraining. [Hrsg.] American Chemical Society. Journal of Chemical Information and Computer Sciences. 1995, Bd. 35, 5, S. 826-833.

Tobler und R., W. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. [Hrsg.] Routledge. Economic Geography. 46, 1970, S. 234-240.

van den Heuvel, Martijn, Mandl, Rene und Hulshoff Pol, Hilleke. 2008. Normalized Cut Group Clustering of Resting-State fMRI Data. [Hrsg.] Public Library of Science. PLOS ONE. 2008, Bd. 3, 4.

Witten, Ian H. und Frank, Eibe. 2001. Data Mining Praktische Werkzeuge und Techniken für das maschinelle Lernen. München Wien : Carl Hanser Verlag, 2001.

Zhang, G. Peter. 2010. Neural Networks For Data Mining. [Buchverf.] Oded Maimon und Lior Rokach. Data Mining and Knowledge Discovery Handbook. 2. New York Dordrecht Heidelberg London : Springer, 2010, S. 419-444.

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, den _____ _____