Mit freundlicher Unterstützung des

ScaDS.AI DRESDEN LEIPZIG

BVM Berufsverband Deutscher Markt- und Sozialforscher e.V.

## BVM Data Science Cup 2020

Seit 2017 führt der BVM jährlich den Data Science Cup durch und bietet damit Marktforschern und interessierten Teilnehmern aus angrenzenden Fachgebieten eine Gelegenheit, ihre Data Science Expertise spielerisch in einem Wettbewerb auf die Probe zu stellen. Die Aufgabenstellungen sind vielfältig und weisen mit Themen wie der Modellierung des Ausrüstungsmarktes eines MMORPG, der Beeinflussung von Wahlergebnissen durch das willkürliche Ziehen von Wahlkreisen („Gerrymandering“) oder eigennützigem Trolling auf Twitter nicht nur einen Marketingbezug sondern auch eine allgemeine gesellschaftliche Relevanz auf. Die beste Lösung wird im Rahmen der Mitgliederversammlung des BVM am 06.11.2020 vorgestellt und prämiert.

Als Partner für den Data Science Cup 2020 konnte der BVM den ADAC gewinnen. Als mitgliederorientierter Mobilitätsdienstleister stellt der ADAC umfangreiche Informationen zur aktuellen Verkehrslage bereit. Dabei ist die genaue Erfassung von Baustellen von besonderem Interesse, da sie die Ursache für eine Vielzahl von Verkehrsbeeinträchtigungen sind. Aktuell stützt sich die Detektion baustellenbedingter Einschränkungen vor allem auf statische Plandaten aus den Baustelleninformationssysteme der Länder. Da diese Daten hinsichtlich Genauigkeit und Aktualität gewissen Grenzen unterliegen, versucht der ADAC das Potential von in Fahrzeugen erhobenen Echtzeitdaten für diesen Anwendungsfall zu erschließen.

### Daten

Die Datenbasis für die diesjährige Aufgabe besteht aus zwei Teilen: Floating Car Data (FCD), die aus fahrenden Fahrzeugen erhoben werden und Plandaten aus den Baustelleninformationssystemen (BIS) der Länder.

Floating Car Data wird über das Staumeldesystem des ADAC und von verschiedenen Flottenbetreibern und Telematik-Dienstleistern erhoben, aufbereitet und für eine Verkehrslagebeurteilung ausgewertet.
Die Daten liegen im CSV-Format vor. Die erste Zeile enthält die Namen der Datenfelder. Die Reihenfolge der Namen legt die Reihenfolge der Datenfelder fest. Die Datensätze folgen ab der zweiten Zeile, wobei jede Zeile eine Positionsmeldung enthält. Alle Namen und Datenfelder werden durch ";" getrennt. Die Werte aller Datenfelder wurden in eine Textdarstellung umgewandelt. Datum- und Zeitfelder werden im Format "dd.mm.yyyy hh:mm:ss" dargestellt, z.B. „12.06.2010 08:34:56“. Datenfelder, die schon im Zeichenketten-Format vorliegen, werden durch doppelte Hochkomma eingeschlossen, z.B. “MÜNCHEN“. Datenfelder im Format „Bool“ werden in die Zeichen „0“ bei FALSE und „1“ bei TRUE kodiert.
Die BIS-Daten liegen ebenfalls als CSV in einem ähnlichen Format vor. Die erste Zeile enthält wieder die Namen der Datenfelder. Die folgenden Zeilen enthalten jeweils eine Baustelle mit den zugehörigen Positions- und Zeitangaben.
Eine detaillierte Auflistung und Beschreibung der jeweiligen Datenfelder befindet sich im Anhang „Datenfelder“.

### Aufgabenstellung

Basierend auf den bereitgestellten Daten, soll ein **Analyseansatz entwickelt werden, der es erlaubt. Baustellenbedingte Verkehrsbeeinträchtigungen zu erkennen**. Auf Grundlage der BIS-Daten gelingt eine solche Erkennung nur unzuverlässig da es häufig zu zeitlichen Verschiebungen bei den Bauarbeiten kommt, die in den Plandaten nicht aktualisiert werden. Eine Baustelle kann in der Realität daher sowohl vor, als auch nach dem geplanten Starttermin beginnen und länger oder kürzer als die geplante Dauer existieren.

Aus diesem Grund liegt der Fokus der diesjährigen Analyseaufgabe auf den FCD-Bewegungsdaten. Diese werden pro Sekunde erfasst und regelmäßig übertragen. Damit bieten sie eine aktuelle Ist-Repräsentation des tatsächlichen Verkehrsflusses. Beeinträchtigungen dieses Verkehrsflusses können also ein Hinweis auf die Existenz einer Baustelle sein.

Mit freundlicher Unterstützung des
ScaDS.AI DRESDEN LEIPZIG
BVM Berufsverband Deutscher Markt- und Sozialforscher e.V.

In diesem Jahr sind die Teilnehmer aufgefordert, einen kompletten Analyseansatz zu entwickeln, zu dokumentieren und anzuwenden. Der Analyseansatz kann frei gestaltet werden, sollte aber die folgenden drei Teile enthalten:

(1.) eine Modellierung für den Verkehrsfluss, bestehend aus Kennzahlen, Metriken etc. die aus den bereitgestellten Daten abgeleitet werden;
(2.) eine Definition bzw. ein Detektionsansatz, der auf dieser Modellierung aufbaut und die Identifikation von Verkehrsbeeinträchtigungen ermöglicht;
(3.) ein Klassifikationsansatz, der entscheidet, ob eine Verkehrsbeeinträchtigung durch eine Baustelle entstanden ist.

Die Dokumentation des Analyseansatzes dient als Grundlage für die Bewertung durch die Jury und soll es ermöglichen, den Ansatz und die mit ihm erzielten Ergebnisse nachzuvollziehen und zu prüfen. Die Dokumentation soll in Form eines Forschungsberichts/Research Papers eingereicht werden und den Ansatz eindeutig, formal und logisch strukturiert beschreiben.

Teilnehmer können für ihren Analyseansatz zusätzliche Datenquellen nutzen, sofern diese öffentlich zugänglich sind und in der Dokumentation referenziert werden.

## Aufgabe 1: Posteriori Bestimmung von Baustellenzeiten

Ziel dieser Aufgabe ist die Bestimmung der tatsächlichen Standzeiten von Baustellen. Die Datenbasis für diese Aufgabe enthält FCD- und BIS-Daten für die Autobahnen des Freistaats Bayern im zweiten und dritten Quartal 2019. Stellen Sie für jede Baustelle die Soll-Standzeit (BIS) und die mit ihrem Ansatz ermittelte Ist-Standzeit gegenüber. Begründen Sie ihre Ergebnisse. Für die Ergebnisse sind alle Baustellen relevant die im betrachteten Zeitraum von März bis September beginnen UND/ODER enden.

Bewertung: Aufgrund der hohen Anzahl an Baustellen wird im Rahmen der Bewertung nur eine Auswahl an Baustellen untersucht. Diese wird in Form eines Excel Sheets bereitgestellt, in das die Teilnehmer ihre Ergebnisse eintragen können. Die Jury vollzieht anhand der Dokumentation die Entstehung der Lösung nach. In Kombination mit den beigefügten Begründungen ergibt sich eine Wertung.

## Aufgabe 2: Vorhersage von Verkehrsbeeinträchtigungen

Nachdem in Aufgabe 1 vom Verkehrsmodell auf die Baustelle geschlossen wurde, sollen in dieser Aufgabe Schlüsse von der Baustelle auf den Verkehrsfluss gezogen werden. Ziel ist es, eine Prognose für ihr Verkehrsflussmodell abzugeben. Die Datenbasis für diese Aufgabe enthält die BIS-Daten für die Autobahnen des Freistaats Bayern im **Februar** 2020. Geben sie für alle Baustellen, die planmäßig in diesem Monat starten UND/ODER enden die Kennzahlen, Metriken etc. an, die ihren Verkehrsfluss (Durchschnittsgeschwindigkeit, Menge an Fahrzeugen etc.) beschreiben pro Stunde an. Sollte eine Baustelle eine Dauer von mehr als einem Tag aufweisen, bilden Sie bitte die durchschnittlichen Stundenwerte pro Wochentag (z.B. Durchschnitt über alles Montage).

Bewertung: Mit den im Februar 2020 erhobenen FCD Bewegungsdaten und der Dokumentation berechnet die Jury die tatsächlichen Werte ihres Verkehrsflussmodells und bestimmt die Abweichungen zu ihrer Prognose. Je geringer die Abweichung desto besser die Wertung.

Mit freundlicher Unterstützung des
ScaDS.AI DRESDEN LEIPZIG
BVM Berufsverband Deutscher Markt- und Sozialforscher e.V.

**Anhang Datenfelder**

| Datenfelder FCD | | |
|---|---|---|
| **Name** | **Typ** | **Bedeutung** |
| oid | string | *optional* |
| providerId | int | *Identifikation der Meldungsquelle / optional* |
| assetid | string | *Identifikation des Fahrzeuges / verschlüsselt* |
| longitude | int | *Longitude der aktuellen GPS-Koordinate * 1000000 (Dezimalgrad) / WGS84* |
| latitude | int | *Latitude der aktuellen GPS-Koordinate * 1000000 (Dezimalgrad) / WGS84* |
| lastLongitude | int | *Longitude vorhergehenden der GPS-Koordinate * 1000000 (Dezimalgrad) / WGS84 / kann leer sein* |
| lastLatitude | int | *Latitude der vorhergehenden GPS-Koordinate * 1000000 (Dezimalgrad) / WGS84 / kann leer sein* |
| status | int | *Leer / undefiniert* |
| velocity | int | *Geschwindigkeit des Fahrzeuges in km/h* |
| direction | int | *GPS-Angabe der Fahrtrichtung des Fahrzeuges in Zehntelgrad / 0=Nord, 900 = OST, 1800 = SÜD, 2700 = WEST* |
| timestamputc | DateTime | *GPS-Zeitstempel des Bordgerätes oder des Providers (UTC)* |
| poshint | text | *Leer* |
| created | DateTime | *Leer* |
| country | string | *Staat als ISO-2-Kode (z.B. de für Deutschland)* |
| federalState | int | *Bundesland (Kodierung)* |
| district | int | *Landkreis (Kodierung)* |
| town | int | *Gemeinde (Kodierung)* |
| isdirectionvalid | bool | *Gültigkeit der Fahrtrichtungsangabe im Feld forwarddir (1=gültig, 0=ungültig)* |
| forwarddir | bool | *1=Fahrzeug fährt in positiver TMC-Richtung, 0=Fahrzeug fährt in negativer TMC-Richtung* |
| vehicletype | int | *Fahrzeugtyp, mögliche Werte: 0=unbekannt 1=LKW 2=Kleintransporter 3=PKW* |
| tmcsegmentid | string | *Datenbank-Referenz auf das TMC-Segment (Straßenabschnitt)* |
| tmc_longitude | int | *GPS-Koordinate * 1000000 Longitude der auf das TMC-Segment abgebildeten Position (Dezimalgrad)* |
| tmc_latitude | int | *GPS-Koordinate * 1000000 Latitude der auf das TMC-Segment abgebildeten Position (Dezimalgrad)* |
| distfrom | int | *Entfernung des Fahrzeuges zum Anfang des Straßensegmentes in Meter* |
| distto | int | *Entfernung des Fahrzeuges zum Ende des Straßensegmentes in Meter* |

Mit freundlicher Unterstützung des
ScaDS.AI DRESDEN LEIPZIG
BVM Berufsverband Deutscher Markt- und Sozialforscher e.V.

| **Datenfelder BIS** | | |
|---|---|---|
| **Name** | **Typ** | **Bedeutung** |
| id | int | *Meldungsnummer* |
| countrycode | string | *Kürzel des Staates der Von-Lokation, hier D* |
| federalcode | string | *Kürzel des Bundeslands der Von-Lokation, hier BY* |
| roadnumber | string | *Nummernbezeichnung für Straße* |
| roadname | string | *Straßenname* |
| region | string | *Lokationsname bei Regionsmeldungen* |
| segmentfrom | string | *Autobahnabschnitt Von* |
| segmentto | string | *Autobahnabschnitt Nach* |
| locationfrom | string | *Name Von-Lokation* |
| locationto | string | *Name Nach-Lokation* |
| tmcfrom | int | *TMC Code Von-Lokation* |
| tmcto | int | *TMC Code Nach-Lokation* |
| viaidfrom | int | *ViaNummer Von-Lokation* |
| viaidto | int | *ViaNummer Nach-Lokation* |
| exitnrfrom | int | *Nummer Anschlussstelle Von-Lokation* |
| exitnrto | int | *Nummer Anschlussstelle Nach-Lokation* |
| tmcdir | bool | *1 = positive TMC Richtung*<br>*0= negative TMC Richtung* |
| xfrom | int | *Longitude (X-Koordinate) Von-Lokation* |
| yfrom | int | *Latitude (Y-Koordinate) Von-Lokation* |
| xto | int | *Longitude (X-Koordinate) Nach-Lokation* |
| yto | int | *Latitude (Y-Koordinate) Nach-Lokation* |
| bothdirection | String | *beide Fahrtrichtungen betroffen (True/False)* |
| tsfromUTC | DateTime | *UTC Zeistempel Start* |
| tstoUTC | DateTime | *UTC Zeistempel Ende* |
| tsfrom | DateTime | *Lokale Zeit Start* |
| tsto | DateTime | *Lokale Zeit Ende* |
| dayfrom | String | *Kürzel des Wochentags für Startdatum* |
| dayto | String | *Kürzel des Wochentags für Enddatum* |
| duration | int | *Dauer in Sekunden* |
| length | int | *Länge in Kilometern* |

Mit freundlicher Unterstützung des
ScaDS.AI DRESDEN LEIPZIG
BVM Berufsverband Deutscher Markt- und Sozialforscher e.V.

| | | |
|---|---|---|
| cause | String | *Ursache: B=Baustelle, D=Defektes Fahrzeug, V=Hohes Verkehrsaufkommen, U=Unfall, X=Sonstiges* |
| effect | String | *Auswirkung auf Verkehr: O=Stockender Verkehr, S=Stau* |
| event01…event12 | String | *Liste mit Zusatzinformationen* |
| text | String | *Zusammenfassung der Baustelleninformation in Textform* |
| source | String | *Kodiert, 1 Zeichen pro Quelle: A=ADAC, P=Polizei, S=Staumelder, E=Eric, X=Panda, L=LKW, R=Rundfunk* |
| loctype | String | *Kodiert, 1 Zeichen pro Typ: F=Fernstraße, R=Region, G=Grenze, S=Stadt, V=Verbindungen, A=Alpenstraße, X=Refcodiert* |
| cat | String | *Kodiert, 1 Zeichen pro Kategorie: V=Verkehrslage, B=Baustelle, P=Panne, Z=Ankündigung, G=Gefahr, W=Wetter, H=Parkplatz, A=Alpenstraße, X=System* |
| ver | int | *Anzahl erstellte Versionen* |
| active | String | *Meldung noch aktiv (True/False)* |
| incidentkey | String | |