

«Mobilitätsbeobachtungssystem» in Pully

Ein Tool zur Quantifizierung des Verkehrs im Stadtzentrum

«Observatoire de la Mobilité» à Pully

Un outil qui permet de quantifier la mobilité dans le centre-ville

Seit 2015 arbeitet die Stadt Pully zusammen mit Swisscom an einem Mobilitätsbeobachtungssystem. Durch Nutzung der zwischen SIM-Karten und Antennen ausgetauschten Signale kann Swisscom Angaben zur Mobilität in klar abgegrenzten Gebieten machen. Dabei werden Ausgangs- und Zielort sowie Aufenthaltsdauer fortlaufend kommuniziert, und herkömmliche, teure Studien sind nicht mehr nötig. Trotzdem gibt es zwei einschränkende Faktoren: die Repräsentativität des Verkehrs und der Umgang mit der Multimodalität. Zur Ergänzung des digitalen Systems wurden physikalische Sensoren installiert. Die von den Sensoren ausgegebenen Daten wurden mit jenen von Swisscom verglichen, um die Qualität von letzteren beurteilen zu können. Wir stellten fest, dass Swisscom den Fahrzeugverkehr viel besser erkennt als Fußgänger, deren Bewegungen komplexer sind. Es wurden mehrere Datenfusionsmodelle in Erwägung gezogen. Alle haben Vorteile, aber auch Nachteile, die weiter untersucht werden müssen.

Seit 2015 setzt die Stadt Pully ein Tool ein, mit dem die Mobilität im Stadtzentrum mittels Analyse von Parametern wie multimodale Verkehrsflüsse oder Geschwindigkeiten quantitativ erfasst werden kann. Dieses System zur Mobilitätsbeobachtung nutzt die Plattform «Mobility Insights» von Swisscom – ein analytisches Tool, das auf die urbane Mobilität angewandt wird^[1]. Das System nutzt die Milliarden von Einzeldaten, die Abonnenten täglich erzeugen: Anrufe, SMS, Internetabfragen und andere Aktivitäten, die eine Verbindung

La ville de Pully a développé depuis 2015 un Observatoire de la Mobilité, en collaboration avec Swisscom. En exploitant les signaux échangés entre cartes SIM et antennes, Swisscom parvient à fournir des informations sur la mobilité dans des zones précises. Les origines/destinations ainsi que les temps de séjour sont communiqués en continu, évitant ainsi les traditionnelles et coûteuses enquêtes. Cela étant, deux limites se présentent: la représentativité du trafic et la gestion de la multimodalité. Une architecture de capteurs physiques a été installée en complément au système digital. Les données issues des capteurs ont été confrontées à celles de Swisscom pour évaluer la qualité de ces dernières. Nous avons remarqué que Swisscom détecte beaucoup mieux les trafics de véhicules que les piétons, qui ont une mobilité plus complexe. Plusieurs modèles de fusion des données ont été proposés, avec chacun des avantages et limites qui nécessitent de plus amples investigations.



PAR

JANODY Pougala

Docteurante dans le laboratoire de Transport et Mobilité (TRANSP-OR) à l'EPFL

VSS-Preis | Prix VSS 2020

Janody Pougala erhielt den VSS-Preis 2020 für ihre Masterarbeit an der ETH Lausanne.

Janody Pougala a obtenu le prix VSS 2020 pour le mémoire de master à l'EPFL.

En 2015, la ville de Pully crée un outil qui permet de quantifier la mobilité dans le centre-ville, en estimant des paramètres tels que des flux multimodaux ou des vitesses. Cet «Observatoire de la Mobilité» utilise la plateforme «Mobility Insights» de Swisscom, un outil à portée analytique appliqué à la mobilité urbaine^[1]. Le système utilise les milliards de données générées quotidiennement par les abonnés: appels, SMS, requêtes Internet, et autres activités qui nécessitent une connexion avec une

zu einer Antenne erfordern. Ein Algorithmus bearbeitet diese Daten und erzeugt daraus Mobilitätsstatistiken. Insbesondere die Karten der Ausgangs- und Zielorte der Fahrten und die Schätzungen der in einer Zone verbrachten Zeit überzeugten Pully. Diese Elemente geben Auskunft zum Einflussbereich des Stadtzentrums (Woher kommen die Besucher?) und zur Dauer der Besuche (Wie lange bleiben sie?). Das System von Swisscom eröffnet auch die Möglichkeit, das Mobilitätsmonitoring, das sich bis anhin auf verschiedene Werkzeuge und Methoden stützte (physische Sensoren, Erhebungen usw.), zentral zu organisieren.

Dennoch ist aufgrund der technischen Grenzen dieses Systems zur Mobilitätsbeobachtung (Genauigkeit der Erkennung und Ortung, Datenfiltrierung usw.) eine Kalibrierung notwendig, um eine gewisse Qualität garantieren zu können. Das ist jedoch nicht möglich ohne einen Vergleich mit vor Ort gewonnenen Daten, die manuell oder automatisch von physischen Zählern erfasst wurden. Außerdem deckt Swisscom nicht den ganzen Markt ab, und das Tool funktioniert nur bei eingeschalteten Smartphones^[2]. Eine Aufbereitung der Daten wird vom Unternehmen durchgeführt – basierend auf den Marktanteilen, die für die Herkunftsgemeinde jeder SIM berechnet wurden –, aber Verzerrungen der Repräsentativität sind dennoch zu beachten. Bei den ersten Durchläufen des Algorithmus wird deshalb kein Modal-split vorgenommen: Die Fahrten werden gemäss der genutzten Infrastruktur – Strasse, Autobahn oder Schiene – eingeteilt. Eine genauere Klassifizierung (Langsamverkehr, motorisierter Individualverkehr MIV oder öffentliche Verkehrsmittel ÖV) ist daher nicht möglich. Die Bedeutung des Tools wird dadurch gemindert, denn es liefert nur Informationen zum Fahrzeugverkehr.

Das Beobachtungssystem birgt ein Potenzial, das – zumindest vorerst – nur mit einer Kombination mehrerer Sensoren voll ausgenutzt werden kann. Dieses Projekt will diesem Bedürfnis gerecht werden, indem es ein System aus physikalischen Sensoren entwickelt und umsetzt und damit Folgendes ermöglicht:

- Informationen liefern zu den multimodalen Verkehrsflüssen im Stadtzentrum, insbesondere die Bewegungen von Fußgängern, über die noch nichts bekannt ist;
- als Versuchsumgebung dienen, um digitale Daten zu kalibrieren.

antenne. Un algorithme traite ces données et en tire des statistiques de mobilité. Notamment, les cartes d'origines et destinations des trajets à échelle de la Suisse et les estimations de temps passé dans une zone ont séduit Pully. Ces éléments permettent d'obtenir des informations telles que le rayonnement géographique de leur centre-ville (d'où viennent les visiteurs?) et des durées de visite (combien de temps restent-ils?). Le système de Swisscom laisse envisager la possibilité de centraliser le monitoring de la mobilité, qui jusqu'alors exigeait une multiplicité d'outils et méthodes (capteurs physiques, enquêtes, etc.)

Cependant, vues les limites techniques de l'Observatoire (précision des détections et localisations, filtrage des données, ...), un calibrage est nécessaire pour garantir une certaine qualité. Or, celui-ci n'est possible que par comparaison avec des données de terrain, recueillies manuellement ou de façon automatique par des compteurs physiques. De plus, Swisscom ne touche pas la totalité du marché, et l'outil ne fonctionne qu'en présence de smartphones allumés^[2]. Un réhaussement des données est effectué par l'entreprise, basé sur les parts de marché calculées pour la commune d'origine de chaque SIM, mais des biais de représentativité sont à prévoir. Enfin, dans les premières itérations de l'algorithme, aucune discrimination des parts modales n'est effectuée: les trajets sont classés selon l'infrastructure empruntée – route, autoroute ou chemins de fer. Il est alors impossible d'avoir une classification plus fine (modes doux, transports individuels motorisés TIM ou transports en commun TC). L'intérêt de l'outil est donc amoindri s'il ne peut fournir d'informations que sur le trafic de véhicules.

L'Observatoire présente un potentiel, qui, pour l'instant du moins, nécessite une combinaison de plusieurs capteurs pour être exploité au maximum. Ce projet propose de répondre à ce besoin, avec le développement et l'implémentation d'une architecture de capteurs physiques, permettant de:

- Renseigner sur les flux multimodaux dans le centre-ville, en particulier les mouvements piétons, jusqu'à présent inconnus.
- Servir de banc d'essai pour calibrer les données digitales.

Anzeige

MAS

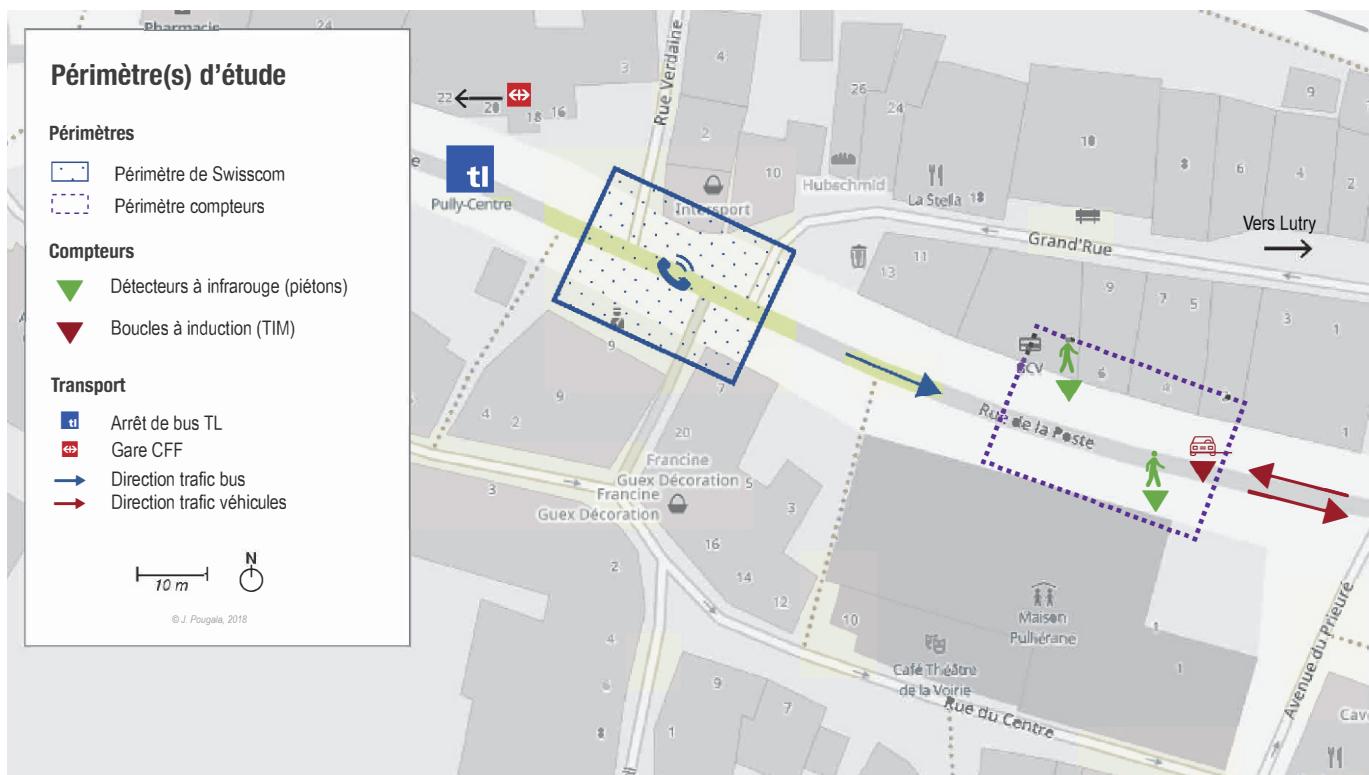
Infrastruktur und Verkehr
Infrastructure et transport
Infrastructure and transport

B
Berner Fachhochschule

VSS <>

Praxisnahe Weiterbildungen – informieren Sie sich jetzt über das Kursangebot 2020–2024

www.vss.ch/aus-und-weiterbildung



1 | Versuchsbereich.
1 | Périmètre d'étude.

Es ist nicht das Ziel, das bestehende System zu ersetzen, sondern eine Methodologie für die Kombination digitaler und physikalischer Daten auszuarbeiten. Darauf wird ein besonderes Augenmerk gelegt. Diese Zusammenführung zweier Quellen macht es möglich, ein vollständiges Bild der multimodalen Mobilität über einen Zeitraum zu erhalten.

Systemarchitektur für die Datenerfassung

Das Gebiet für den Aufbau der ergänzenden Erfassungsarchitektur ist ein Bereich von ungefähr 20 m x 40 m (Abb.1). Wir beschlossen, das Gebiet so klein wie möglich zu halten, um die Belastung zu beschränken und eine Kalibrierung vorzunehmen.

Vor Beginn des Projekts wurde eine Untersuchung der bestehenden Technologien durchgeführt. Mit einer auf mehreren Kriterien basierenden Analyse wurden die besten Optionen für Pully in Bezug auf Preis, Installation und Unterhalt, Robustheit, Messqualität, Umgang mit der Multimodalität, Integration in die Umgebung und Schutz der Privatsphäre festgelegt. Die Versuchsanordnung setzte sich schliesslich wie folgt zusammen:

- zwei Induktionsschleifen zum Zählen des motorisierten Verkehrs (MIV),
- zwei passive Infrarotsensoren zum Zählen der Fußgänger,
- Nutzungsstatistiken für die zwei Buslinien, die durch das Gebiet führen, die von den Verkehrsbetrieben (Transports publics de la région lausannoise TL) zur Verfügung gestellt wurden.

L'objectif n'étant pas de supplanter le système existant, une attention particulière a été portée à la formalisation d'une méthodologie de combinaison des données digitales et physiques. Cette fusion des deux sources permettrait d'aboutir à une image complète de la mobilité multimodale en fonction du temps.

Architecture pour la collecte des données

Le périmètre choisi pour l'installation de l'architecture complémentaire est une zone d'environ 20 m x 40 m (Figure 1). Nous avons choisi ici de réduire la zone au maximum afin de limiter les nuisances et procéder à un calibrage.

Une revue des technologies existantes a été réalisée en amont du projet, et une analyse multicritère a permis d'identifier les meilleures options pour la ville de Pully en termes de prix, installation et entretien, durabilité, qualité des mesures, gestion de la multimodalité, intégration dans le paysage et respect de la vie privée. Le banc d'essai final se compose de:

- deux boucles à induction pour le comptage du trafic de véhicules motorisés (TIM),
- deux détecteurs à infrarouge passif pour le comptage de piétons,
- des statistiques de fréquentation concernant les deux lignes de bus traversant la zone, fournies par l'opérateur de transports publics (Transports publics de la région lausannoise - TL).

Die Zusammenführung dieser Quellen ist kein leichtes Unterfangen und muss so geschehen, dass ein besseres Ergebnis erzielt wird, als es der Fall wäre, wenn die Quellen einzeln ausgewertet würden. Es ist daher sinnvoll, sich drei vorbereitende Fragen zu stellen:

- Sind alle Quellen zweckmässig?
- Sind alle Quellen (in Bezug aufeinander) kohärent?
- Sind alle Quellen miteinander kompatibel?

Sind die Quellen zweckmässig?

Die Zweckmässigkeit der Datenquellen wird in Bezug auf Korrektheit und Genauigkeit bestimmt: Jeder Sensor muss seine Nutzer zuverlässig erkennen. Daher führten wir manuelle Zählungen durch, um die Zähler zu kalibrieren. Die Schleifen sind relativ leistungsstark (in der Grössenordnung von höchstens rund 20 % relative Abweichung bei den Beobachtungen zur Hauptverkehrszeit), und die Fussgängerströme werden oft unterschätzt, insbesondere wenn sich diese in Gruppen bewegen. Es wurde eine Kalibrierung dieser Messungen durchgeführt. Daraus entwickelten wir anschliessend eine Methodologie, um die Korrektur der gemessenen Ströme zu systematisieren, indem wir die Wahrscheinlichkeit der Beobachtung von Gruppen in einem bestimmten Zeitraum berücksichtigen.

Beim Einsatz physikalischer Zähler ist es wichtig, bei der Installation abzuschätzen, wie repräsentativ sie für den tatsächlichen Verkehr sind. Da die Messungen dieser Sensoren punktuell sind, kann zwangsläufig lediglich ein Teil des Verkehrsaufkommens erfasst werden. Das kann zu Divergenzen mit den Hochrechnungen von Swisscom führen, die flächenbezogen und deshalb dieser Einschränkung weniger stark unterworfen sind. Für die Induktionsschleifen stellte sich dieses Problem nicht: Die betreffende Strasse ist die einzige, die unseren Versuchsbereich durchquert, daher führt der gesamte Fahrzeugverkehr zwingend am Messpunkt vorbei. Im Gegensatz dazu war der Fussgängerstrom nicht immer gleich repräsentativ: Das Vorhandensein von Trottoirs garantiert nicht, dass diese auch benutzt werden – ein Phänomen, das dadurch verstärkt wird, dass die untersuchte Zone nur wenig befahren ist. Das Vorhandensein von Parkplätzen, Geschäften, Haltestellen und anderen Anziehungspunkten für Fussgängerbewegungen auf dem Gebiet vervielfacht das Risiko «chaotischer» Bewegungen, die nicht automatisch gemessen werden können. Wir schätzten die Erfassungsgenauigkeit der Fussgängerzähler auf rund 40 % bis 60 % des tatsächlichen Fussgängerstroms an Werktagen und auf bis 80 % in verkehrsarmen Zeiten oder am Wochenende.

Sind die Quellen (in Bezug aufeinander) kohärent?

Sobald die Qualität der Datenquellen sichergestellt war, wollten wir die von Swisscom erhaltenen Hochrechnungen (ohne Unterscheidung der verschiedenen Verkehrsmodi) mit den Messungen des ergänzenden Systems vergleichen. Um das Zusammenspiel der Daten simultan zu nutzen, wurden die Fahrzeugflüsse anhand der durchschnittlichen Belegungsrate $f_{bel} = 1,1$ in Nutzerflüsse umgerechnet. Wir trafen auch die Annahme, dass jede Person nur eine SIM-Karte mit sich führt.

La fusion de ces sources de données n'est pas anodine, et doit être effectuée de façon à assurer que le résultat soit meilleur que dans le cas où les sources sont considérées individuellement. Pour cela, il convient de se poser trois questions préliminaires:

- Toutes les sources sont-elles pertinentes?
- Toutes les sources sont-elles cohérentes (entre elles)?
- Toutes les sources sont-elles compatibles?

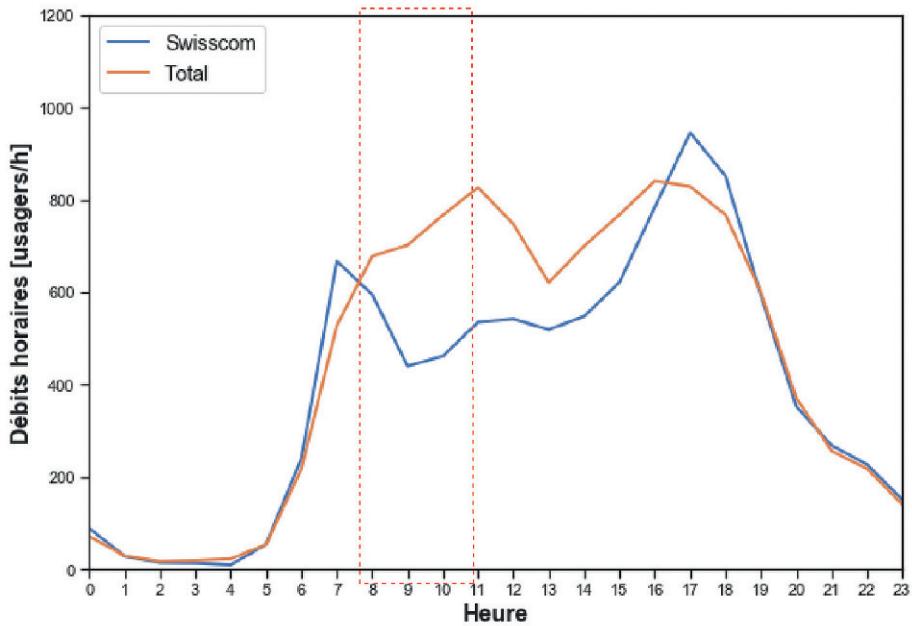
Les sources sont-elles pertinentes?

La pertinence des sources de données est définie en termes d'exactitude et de précision: chaque capteur doit détecter ses usagers de façon fiable. Pour cela, nous avons effectué des séances de comptage manuel afin de calibrer les compteurs. Si les boucles sont relativement performantes (autour de 20 % d'écart relatif maximal avec les observations en heure de pointe), les flux de piétons sont souvent sous-estimés, en particulier les cas de déplacements en groupe. Un calibrage de ces mesures a été effectué; et nous avons abouti à une méthodologie pour formaliser la correction des débits mesurés, en considérant la probabilité d'observer des groupes pendant une période donnée.

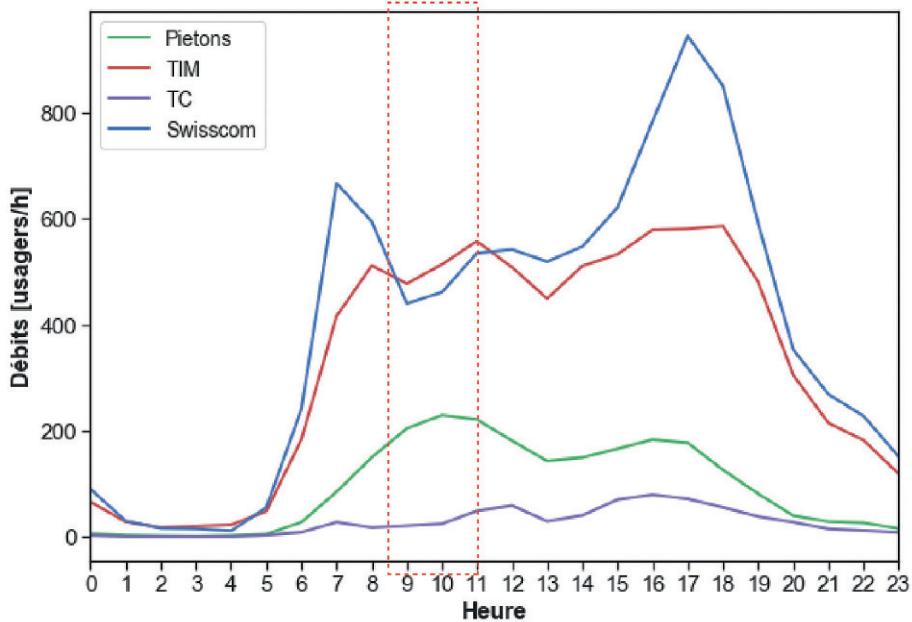
Il est important, lors de la mise en place de compteurs physiques, de prendre en considération leur représentativité du trafic réel lors de leur installation. Étant donné que les mesures de ces capteurs sont ponctuelles, il est inévitable qu'une partie du trafic ne puisse être mesurée. Ceci peut engendrer des divergences avec les estimations de Swisscom, qui elles sont surfaciques et sont donc moins exposées à cette limite. Ce problème ne s'est pas posé pour les boucles à induction: la route concernée est la seule traversant notre périmètre d'étude, contraignant ainsi l'intégralité du trafic de véhicules à passer par le point de mesure. A l'inverse, la représentativité des flux de piétons est très variable: la présence de trottoirs ne garantit pas qu'ils seront utilisés – et ceci est d'autant plus vrai que la zone étudiée est calme en termes de trafic de véhicules. La présence dans la zone d'éléments tels que des places de stationnement, des commerces, arrêts de transports ou autre pôles attracteurs de déplacements piétons, multiplie le risque de déplacements «anarchiques» – et donc, impossibles à mesurer de façon automatique. Nous avons estimé une représentativité des compteurs piétons de l'ordre de 40 à 60 % du débit réel de piétons, en période ouvrable, et jusqu'à 80 % en périodes creuses et le week-end.

Les sources sont-elles cohérentes?

Une fois la qualité des sources de données vérifiées, nous avons souhaité comparer les estimations fournies d'un côté par Swisscom (sans distinction des différents modes), avec les mesures de l'architecture complémentaire. Afin d'exploiter les jeux de données simultanément, les débits véhicules sont convertis en débits d'usagers en les multipliant par un taux d'occupation moyen $f_{occ} = 1,1$. On fait également l'hypothèse que chaque personne n'est équipée que d'une carte SIM.



2 | Durchschnittliche stündliche Flüsse.
2 | Débits horaires moyens.



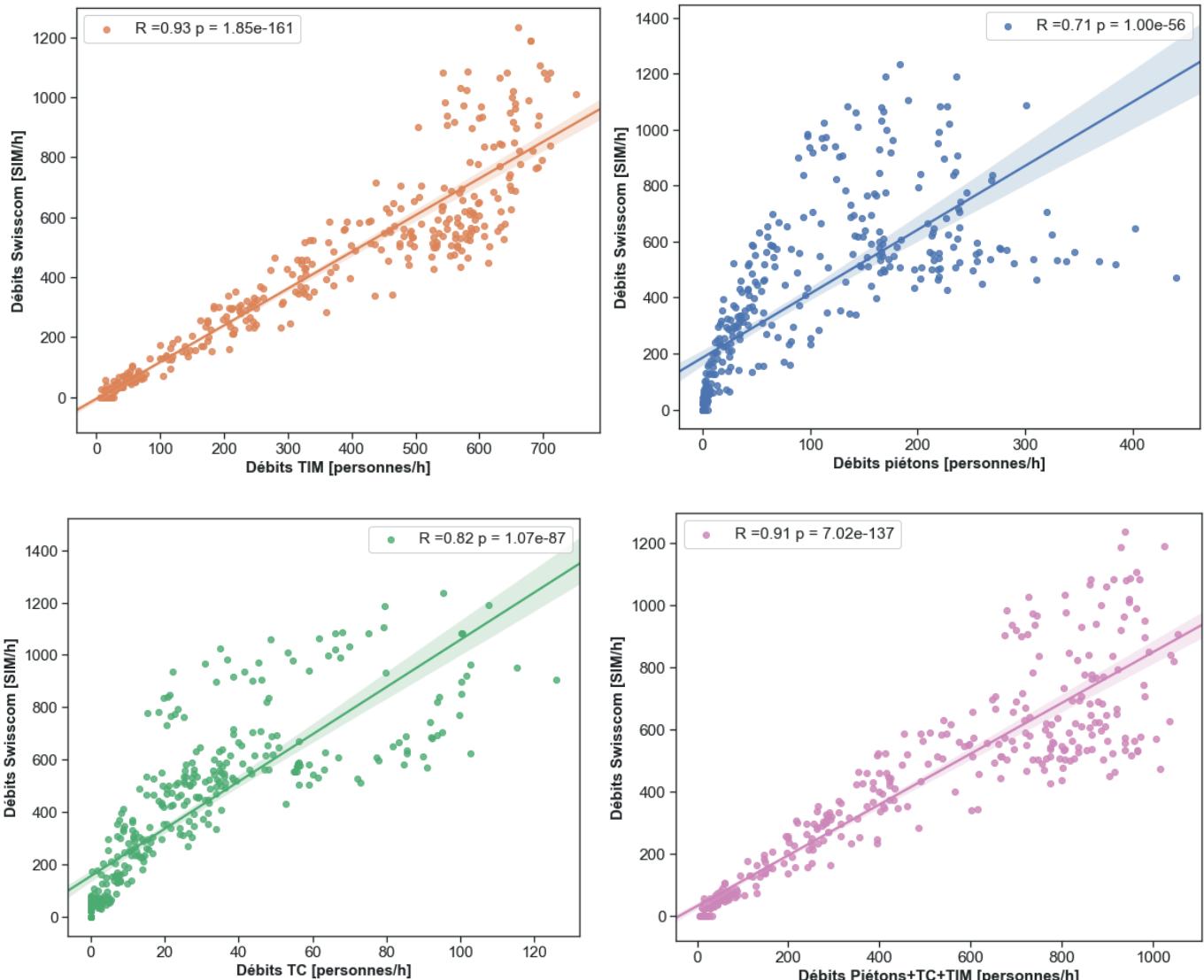
3 | Durchschnittliche stündliche Flüsse
(alle Verkehrsmodi).
3 | Débits horaires moyens (tous modes).

Die Analyse der durchschnittlichen Flüsse pro Stunde (Abb. 2) zeigte uns, dass die Quellen im Allgemeinen kohärent sind. Sie zeigte aber auch nicht vernachlässigbare Abweichungen auf. Während die Kurven in verkehrsarmen Zeiten praktisch identisch sind, scheint sich die Genauigkeit der Swisscom-Daten zwischen 7 Uhr und 19 Uhr zu verschlechtern. Die festgestellten Flüsse sind im Vergleich zu den Zählern zwischen 9 Uhr und 11 Uhr besonders gering, was offenbar auch die Zeit ist, in der sich am meisten Fußgänger in der Zone befinden. Das von Swisscom erzeugte Tagesprofil scheint mit dem Profil des motorisierten Verkehrs (Abb. 3) besser übereinzustimmen, obwohl jenes der Fußgänger weniger eindeutige Spitzen zeigt, die zudem zu unterschiedlichen Zeiten auftreten.

Eine eingehendere Untersuchung der Korrelationen (Abb. 4) bestätigt die Hypothese: Die Beziehung zwischen den Fahrzeugflüssen und den Erfassungen von Swisscom ist praktisch

L'analyse des débits horaires moyens (Figure 2) permet de nous rendre compte que les sources sont généralement cohérentes. Elle met cependant en évidence des écarts non négligeables. En effet, si les trafics pendant les heures creuses sont quasiment identiques, entre 7h et 19h, la performance de Swisscom semble se dégrader. Les débits détectés sont particulièrement faibles comparés aux compteurs entre 9h et 11h, ce qui semble correspondre au moment où les piétons sont le plus présents dans la zone. Le profil généré par Swisscom sur la journée semble plus en adéquation avec le profil des trafics véhicules (Figure 3), alors que celui des piétons présente des pointes plus diffuses et à des moments différents.

Une étude plus poussée des corrélations (Figure 4) confirme cette hypothèse: la relation entre les débits de véhicules et les detections de Swisscom est quasi-linéaire. En revanche,



4 | Korrelationen zwischen verschiedenen Quellen.

4 | Corrélations entre différentes sources.

linear. Bei der Beziehung zu den Fußgängerströmen ist dies hingegen nicht der Fall: Der Graph zeigt eine starke Verteilung, wenn die Fußgängerflüsse 100 Personen/h überschreiten. Die Korrelation mit der Gesamtheit der Sensoren lässt ebenfalls auf eine lineare Beziehung schließen. Der Grund dafür ist das Vorherrschen von Fahrzeugen im Gesamtverkehr (durchschnittlich 75 % aller Verkehrsteilnehmer).

Sind die Quellen miteinander kompatibel?

Die Kompatibilität der Quellen ist eine entscheidende Voraussetzung, um die Fusion der Daten zu ermöglichen. Die Quellen unterscheiden sich durch:

- die **Nutzer, die sie erkennen, und die Bevölkerung, die sie repräsentieren** (z. B. die Erkennungen von Swisscom müssen über deren Abonnenten hinaus hochgerechnet werden);
- die **räumliche Ausdehnung**: Die physikalischen Sensoren führen nur punktuelle Erkennungen durch, während die Erkennungen von Swisscom flächenbezogen sind;
- die **Häufigkeit der Erhebung und der minimalen Zusammenfassung** von Daten (die Swisscom-Daten werden bei-

ce n'est pas le cas de la relation avec les débits piétons: le graphe présente une forte dispersion lorsque les débits piétons dépassent les 100 personnes/h. La corrélation avec la somme des capteurs indique aussi une relation linéaire. Ceci est dû à la prévalence des véhicules dans le trafic total (en moyenne 75 % du trafic).

Les sources sont-elles compatibles?

La question de la compatibilité des sources est primordiale pour être en mesure de fusionner les données. Les sources se différencient par:

- les **usagers qu'elles détectent, ainsi que la population qu'elles représentent** (e.g. les détections de Swisscom doivent être réhaussées au-delà de leurs abonnés);
- leur **portée spatiale**: les capteurs physiques ne font que des détections ponctuelles alors que les détections de Swisscom sont surfaciques;
- la **fréquence de collecte et l'agrégation minimale** de données (e.g. les données Swisscom sont agrégées par heure, alors que les capteurs piétons fournissent des statistiques pour des périodes de 15 minutes);

- spielsweise stündlich zusammengefasst, während die Fußgängerzähler die Statistiken jeweils für einen Zeitraum von 15 Minuten ausgeben);
- die (physischen) **Einschränkungen**, denen sie unterliegen: Strassengraphen, vordefinierte Zonen usw.

Diese Unterschiede erfordern es, dass eine gemeinsame Darstellungsform gewählt wird. Wir haben uns für stündliche Flüsse als Bezugspunkte zwischen den verschiedenen Quellen entschieden.

Modelle für die Datenfusion

Für die Datenfusion zogen wir mehrere Modelle in Betracht:

1. Eine lineare Regression
2. Ein analytisches Modell (nicht lineare Optimierung)
3. Ein Wahrscheinlichkeitsmodell

Aus Platzgründen wird hier nur das Wahrscheinlichkeitsmodell besprochen, da dieses die besten Ergebnisse lieferte.

Wahrscheinlichkeitsmodell

Angesichts der vielen Unsicherheiten der Ausgangslage kann es sinnvoll sein, einen nicht-deterministischen Ansatz zu wählen. Es wird angenommen, dass jeder Verkehrsmodus gemäß einem bekannten Verteilungsgesetz als Zufallsvariable modelliert werden kann. Abbildung 5 zeigt die Verteilung der drei Verkehrstypen im Tagesverlauf, gemessen von den Zählern und den ÖV-Statistiken. Es ist festzustellen, dass jeder Modus charakteristische Bereiche der stündlichen Flüsse abdeckt: Die Flüsse des MIV weisen tendenziell Werte um 500 Fahrzeuge/h (Spitzenbeobachtungen) auf. Die Fußgängerströme belaufen sich hingegen auf rund 200 Fußgänger/h. Diese zwei Verkehrsmodi zeigen auch erhöhte Häufigkeiten für schwächere Flüsse (zwischen 0 und 100 Nutzer/h), was den verkehrsarmen Zeiten entspricht. Die Flüsse der ÖV-Nutzer zeigen ein einheitlicheres Profil mit rund 10 Passagieren/h.

- les **contraintes** (physiques) qui leur sont imposées: graphes routiers, périmètres prédéfinis ...

Vues ces différences il est nécessaire d'aboutir à un mode de représentation commun. Nous avons choisi d'utiliser les débits horaires comme point de référence entre les différentes sources.

Modèles de fusion de données

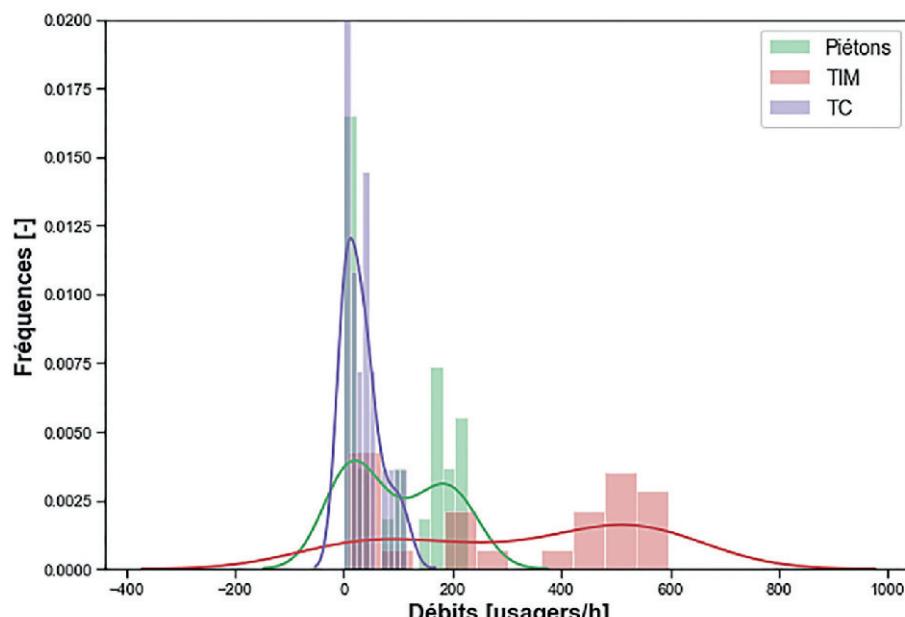
Nous avons proposé plusieurs modèles de fusion des données:

1. Une régression linéaire
2. Un modèle analytique (optimisation non-linéaire)
3. Un modèle probabiliste

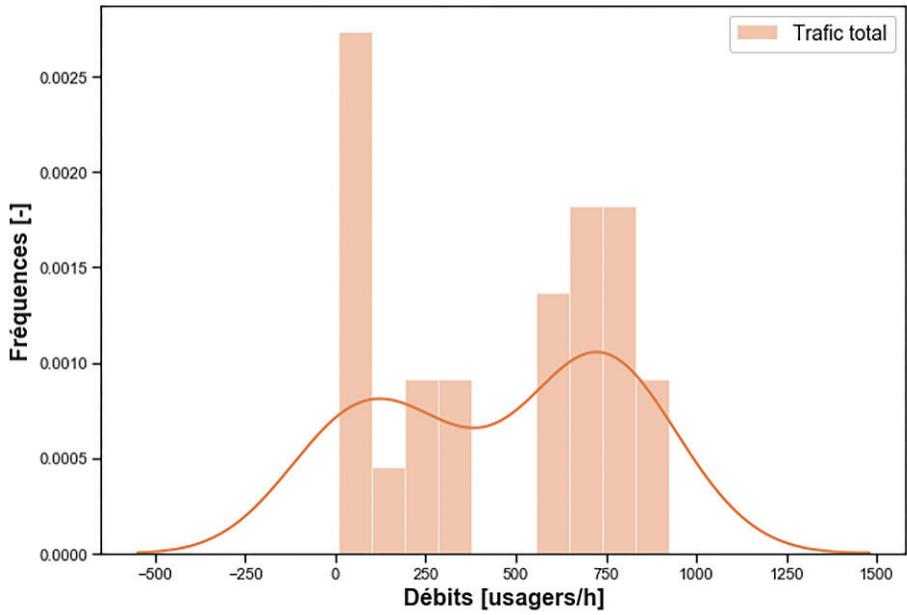
Par souci de brevité, seul le modèle probabiliste, qui a fourni les meilleurs résultats, est présenté dans cet article.

Modèle probabiliste

Étant donné le nombre d'incertitudes du problème, il peut être judicieux d'adopter une approche non déterministe. On considère que chaque mode de transport peut être modélisé comme une variable aléatoire suivant une loi de distribution connue. La Figure 5 montre les distributions journalières des trois types de trafic, mesurés par les compteurs et les statistiques TC. On remarque que chaque mode occupe des plages caractéristiques de débits horaires: les débits de TIM ont tendance à avoir des valeurs qui se regroupent autour de 500 véh/h (observations de pointe), les piétons ont plutôt des débits autour de 200 piétons/h. Ces deux modes présentent également des fréquences élevées pour des débits plus faibles (entre 0 et 100 usagers/h), correspondant aux heures creuses. Les débits des usagers de TC ont un profil plus uniforme, autour de 10 passagers/h.



5 | Verteilung der stündlichen Flüsse.
5 | Distribution des débits horaires.



6 | Gesamtverteilung der stündlichen Flüsse.
6 | Distribution des débits horaires totaux.

Die Verteilungen jeden Modus – mit Ausnahme des ÖV – sind komplex, da sie sich offenbar aus zwei Gauss-Kurven zusammensetzen. Um das Modell zu vereinfachen, wird angenommen, dass alle Modi (Fussgänger F, Autos A und Busse B) einer klassischen Normalverteilung mit einem Mittelwert μ und einer Standardabweichung σ folgen:

$$\begin{aligned} F &\sim N(\mu_f, \sigma_f^2) \\ A &\sim N(\mu_a, \sigma_a^2) \\ B &\sim N(\mu_b, \sigma_b^2) \end{aligned}$$

Ähnliche Muster sind ersichtlich, wenn die drei Modi zusammen betrachtet werden (Abb. 6). Grundsätzlich verteilt sich der Gesamtverkehr auf zwei Flussbereiche: eine erste Spanne mit Flüssen zwischen 0 und 400 Nutzer/h und eine zweite Spanne zwischen 550 und 950 Nutzer/h. Da der Gesamtverkehr aus den drei Verkehrsmodi zusammengesetzt ist, die ihrerseits Normalverteilungen folgen, kann man annehmen, dass das hier geltende Verteilungsgesetz eine Mischung aus diesen Verteilungen ist. Dieses Modell wird als «mixture model» oder Mischmodell bezeichnet. Die Verteilungsfunktion eines Mischmodells mit N Komponenten wird wie folgt definiert^[3]:

$$f(x) = \sum_{i=1}^N \omega_i f_i(x)$$

Wobei:

ω_i : Gewichtung der Mischung der Komponente i ; unter Annahme von $\sum_i \omega_i = 1$

$f_i(x)$: Verteilungsfunktion der Komponente i

Mischmodelle haben den Vorteil, dass sie Soft Clustering (weiche Klassifizierung) ermöglichen, das heißt, für jede Beobachtung kann eine Wahrscheinlichkeit der Zugehörigkeit zu einer Klasse berechnet werden. Die Modelle sind daher flexibler^[4], was das Zusammenführen verschiedener Datenquellen in einem System

Les distributions de chaque mode, à l'exception des TC, sont complexes puisqu'elles semblent composées de deux gaussiennes. Par souci de simplification du modèle, on considérera que tous les modes (Piétons P, Voitures V et Bus B) suivent une loi normale classique, avec moyenne μ et écart-type σ :

$$\begin{aligned} P &\sim N(\mu_p, \sigma_p^2) \\ V &\sim N(\mu_v, \sigma_v^2) \\ B &\sim N(\mu_b, \sigma_b^2) \end{aligned}$$

On retrouve des motifs similaires dans la distribution de leur somme (Figure 6). En effet, le trafic total est distribué autour de deux plages de débits: une première fourchette avec des débits entre 0 et 400 usagers/h, et la deuxième entre 550 et 950 usagers/h. Étant donné que le trafic total dérive des trafics des 3 modes, qui eux suivent des distributions normales, on peut faire l'hypothèse que la loi de distribution qu'il suit est un mélange de ces distributions, appelé «mixture model» ou modèle de mélange en français.

La fonction de répartition d'un modèle de mélange à N composantes se définit de la façon suivante^[3]:

$$f(x) = \sum_{i=1}^N \omega_i f_i(x)$$

Avec:

ω_i : poids de mélange de la composante i ; respectant $\sum_i \omega_i = 1$
 $f_i(x)$: fonction de répartition de la composante i

Les modèles de mélange offrent l'avantage de pouvoir procéder à un soft clustering (classification douce), c'est-à-dire de calculer une probabilité d'appartenance à une classe pour chaque observation. Les modèles sont alors plus flexibles^[4], ce qui facilite l'intégration de différentes sources de données dans un système comportant de nombreuses incertitudes.

mit zahlreichen Unsicherheiten erleichtert. Konkret bedeutet das, dass für jede Beobachtung x des Gesamtverkehrs eine Wahrscheinlichkeit der Zugehörigkeit zu einer der drei Normalverteilungen berechnet wird. Diese Wahrscheinlichkeiten entsprechen den Verteilungsgewichten ω . Die Haupthypothese ist, dass diese Gewichte – bei grossen Stichproben – als modale Teile interpretiert werden können.

Die zweite Hypothese ist, dass die Verteilungsparameter (Mittelwert und Standardabweichung) keinen statistisch bedeutenden Unterschied zwischen den verschiedenen Verkehrsarten zeigen (gezählter Verkehr, tatsächlicher Verkehr, Schätzung von Swisscom). Es wird hier nämlich angenommen, dass dies charakteristische Parameter des Modus und nicht der Beobachtung sind, ebenso wie die Geschwindigkeit beispielsweise.

Die Auflösung erfolgt in zwei Schritten (Abb. 7):

- in einem ersten Schritt, in dem die Parameter basierend auf dem gezählten Verkehr geschätzt werden können; und
- in einem zweiten Schritt, der diese ersten Schätzungen mit den Daten von Swisscom aktualisiert.

Es wird erwartet, dass nach zahlreichen Durchläufen des Modells ein Konvergieren zu der tatsächlichen Verteilung der Verkehrsmodi hin geschieht.

Erster Schritt: Schätzung basierend auf dem gezählten Verkehr
 Die Parameter des Mischmodells werden mithilfe des Erwartungs-Maximierungs-Algorithmus (EM-Algorithmus)^[5, 6] geschätzt. Bei jedem Durchlauf werden Parameter θ für alle Komponenten des Mischmodells erzeugt (in diesem Fall handelt es sich um die Mittelwerte, die Standardabweichungen und die Gewichtungen jeder Unterverteilung, das heisst $\theta_i = \{\mu_i, \sigma_i^2, \omega_i\}$). Die Ausgangswerte dieser Parameter ergeben sich dabei aus den Daten jedes Sensors. Im Erwartungsschritt wird für die gegebenen Beobachtungen und die vorliegenden Parameter die Wahrscheinlichkeit der Zugehörigkeit zu einer der Unterverteilungen i berechnet, das heisst die A-posteriori-Wahrscheinlichkeit $P(x \in i | i)$. Diese Wahrscheinlichkeit wird vom Satz von Bayes abgeleitet.

Anschliessend folgt der Schritt der Maximierung, bei dem die Parameter so neu berechnet werden, dass die Plausibilitätsfunktion maximiert wird, das heisst die Wahrscheinlichkeit, eine Beobachtung x zu erhalten, wenn ein Zusammenwirken der

Concrètement, ceci signifie que pour chaque observation x du trafic total, on calcule la probabilité d'appartenir à une des trois distributions normales. Ces probabilités sont équivalentes au poids des distributions ω . L'hypothèse principale est que, pour de grands échantillons, ces poids peuvent être interprétés comme des parts modales.

La deuxième hypothèse est que les paramètres des distributions (moyenne et écart-type) n'ont pas de différence statistiquement significative d'un trafic à l'autre (trafic compté, trafic réel, estimation de Swisscom). En effet, on considère ici que ce sont des paramètres caractéristiques du mode et non de l'observation, au même titre que la vitesse par exemple.

La résolution se fait en deux étapes (Figure 7):

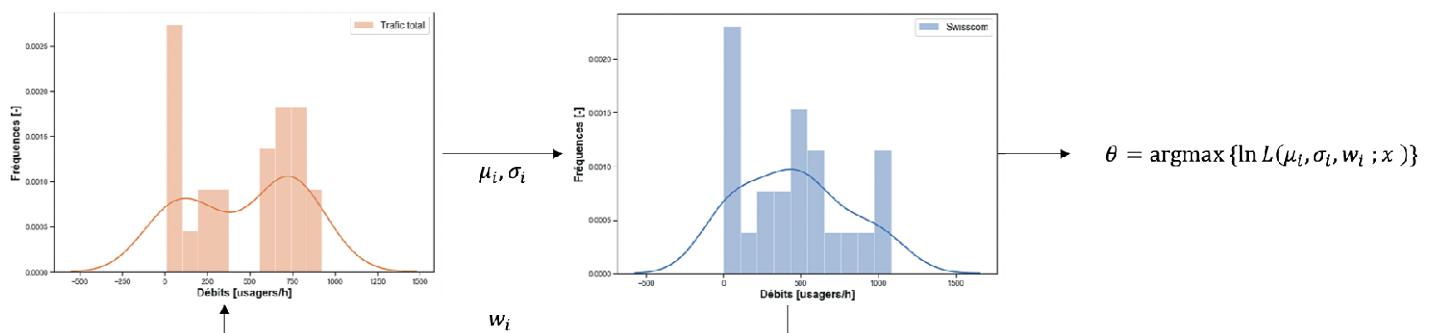
- une première étape qui permet d'estimer les paramètres à partir du trafic compté;
- une deuxième étape qui met à jour ces premières estimations avec les données de Swisscom.

Après de nombreuses itérations de ce modèle, on espère converger vers les distributions réelles des modes.

Première étape: estimation à partir du trafic compté

Les paramètres du modèle de mélange sont estimés à l'aide de l'algorithme d'Espérance-Maximisation (EM)^[5, 6]. A chaque itération, des paramètres θ sont générés pour toutes les composantes du modèle de mélange (dans ce cas, il s'agit des moyennes, écarts-types et poids de chaque sous-distribution, soit $\theta_i = \{\mu_i, \sigma_i^2, \omega_i\}$). Les valeurs initiales de ces paramètres sont ici obtenues à partir des données de chaque capteur. L'étape d'espérance consiste à calculer, pour des observations données et les paramètres fournis, la probabilité qu'elles appartiennent à une des sous-distributions i , c'est-à-dire la probabilité à posteriori $P(x \in i | i)$. Cette probabilité est dérivée de la loi de Bayes.

Ensuite, on procède à l'étape de maximisation, qui consiste à recalculer les paramètres de façon à maximiser la fonction de vraisemblance, c-à-d. la probabilité d'obtenir une observation x étant donné un jeu de paramètres θ . Les estimateurs de vraisemblance de chaque paramètre sont des fonctions de la



7 | Funktionsprinzip des Modells.

7 | Principe du modèle

Parameter θ vorliegt. Die Schätzfunktionen für die Plausibilität jedes Parameters sind die Funktionen der A-posteriori-Wahrscheinlichkeit. Dann wird ein neuer Durchlauf des Algorithmus durchgeführt, wobei die Werte, die sich nach der Maximierung ergaben, als Ausgangspunkte genommen werden.

Das gewählte Endkriterium ist eine Konvergenz des Logarithmus der Plausibilitätsfunktion, das heißt, wenn die Differenz zwischen den erhaltenen Plausibilitäten bei zwei aufeinanderfolgenden Durchläufen unter einem gegebenen Grenzwert liegt.

Zweiter Schritt: Anpassung anhand der Beobachtungen von Swisscom

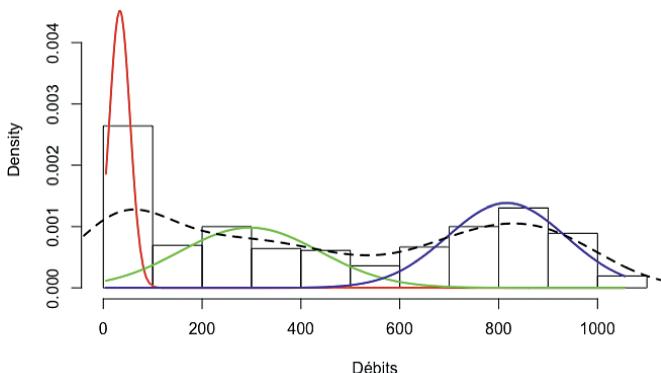
Wenn das Mischmodell in Bezug auf den gezählten Verkehr dimensioniert wurde, wird zum zweiten Schritt übergegangen, bei dem die Parameter an die Swisscom-Daten angepasst werden. Dafür wird ein neuer Erwartungs-Maximierungs-Algorithmus benutzt, wobei aber die Endparameter aus Schritt 1 als Ausgangspunkte genommen werden. Damit kann die Gewichtung für jede Unterverteilung neu berechnet werden.

Ergebnisse

Der Algorithmus wurde in R umgesetzt. Das Modell wurde auf alle stündlichen Flüsse zwischen dem 12. und 26. September 2018 angewandt.

Der erste Schritt, bei dem das Mischmodell angepasst werden konnte, konvergierte in 50 Durchläufen des EM-Algorithmus (Abb. 8). Die drei erhaltenen Unterverteilungen können mithilfe ihres Profils den drei Verkehrsmodi zugeordnet werden. Die erste Verteilung (in der Abbildung rot) entspricht der Verteilung der Busse, die zweite den Fußgängern und die dritte den Fahrzeugen gemäß den jeweiligen Gewichtungen von 0,24, 0,347 und 0,415.

Der zweite Schritt konvergierte in 18 Durchläufen des Erwartungs-Maximierungs-Algorithmus (Abb. 9). Es ist festzustellen, dass das Modell weniger gut eingestellt war, als das beim gezählten Verkehr der Fall war. Tatsächlich scheint ein Teil der Flüsse (insbesondere ein Spitzenwert zwischen 500 und 600 Nutzer/h) mit keiner der Unterverteilungen übereinzustimmen. Ein zwei-



8 | Ergebnis von Schritt 1.
8 | Résultat de l'étape 1.

probabilité à posteriori. On procède alors une nouvelle itération de l'algorithme en utilisant les valeurs obtenues suite à la maximisation comme points de départs.

Le critère d'arrêt choisi est une convergence du logarithme de la fonction de vraisemblance, c-à-d, lorsque la différence entre les vraisemblances obtenues lors de deux itérations successives est inférieure à un seuil donné.

Deuxième étape: ajustement avec les observations de Swisscom

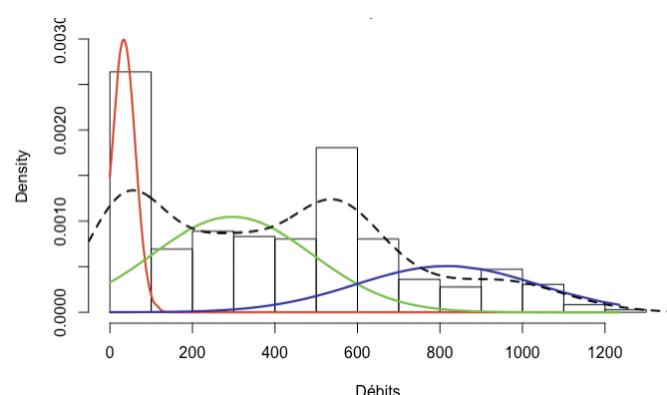
Une fois le modèle de mélange dimensionné par rapport au trafic compté, on procède à la deuxième étape, qui consiste à ajuster les paramètres aux données de Swisscom. Pour cela, on utilise à nouveau l'algorithme d'espérance-maximisation, mais en employant les paramètres finaux de l'étape 1 comme points de départs. Ceci permet de recalculer des poids pour chaque sous-distribution.

Résultats

L'algorithme a été implémenté en R. Le modèle a été appliqué sur tous les débits horaires entre le 12 et le 26 septembre 2018.

La première étape permettant d'ajuster le modèle de mélange a convergé en 50 itérations de l'algorithme d'EM (Figure 8). Les 3 sous-distributions obtenues peuvent être identifiées à des modes à l'aide de leur profil. La première distribution (en rouge sur la figure), s'apparente à la distribution des bus, la deuxième aux piétons et la troisième aux véhicules, pour des poids respectifs de 0,24, 0,347 et 0,415.

La deuxième étape a convergé en 18 itérations de l'algorithme d'espérance maximisation (Figure 9). On remarque que le modèle est moins bien ajusté qu'avec les trafics comptés. En effet, une partie des débits (particulièrement un pic entre 500 et 600 usagers/h) ne semble coïncider avec aucune des sous-distributions. Une deuxième itération des deux étapes a été effectuée, mais le critère d'arrêt a été atteint (pas de changement de la fonction de vraisemblance).



9 | Ergebnis von Schritt 2.
9 | Résultat de l'étape 2.

ter Durchlauf des zweiten Schritts wurde ausgeführt, aber das Endkriterium wurde erreicht (keine Veränderung der Plausibilitätsfunktion).

Bemerkungen

Die Haupthypothese des Modells ist, dass die Variablen den bekannten Verteilungsgesetzen folgen. Trotzdem ist die Vereinfachung auf die Normalverteilung zu unpräzise, um die Flüsse korrekt vorhersagen zu können. Es könnte ein Verteilungsgesetz gewählt werden, das besser an die Modellierung von Verkehr angepasst ist, etwa die Poisson-Verteilung, auf null trunkiert, was den Vorteil hätte, dass negative Werte nicht zugelassen wären.

Rauschen und Messfehler wurden bei der Konzeption des Modells vernachlässigt, aber beim Betrachten der Ergebnisse wird klar, dass es schwierig wäre, in den Daten vorhandene Anomalien zu erkennen. Idealerweise müsste keine Anzahl Verteilungen vorgegeben werden, wie das hier der Fall war, wo drei Unterverteilungen festgelegt wurden. Vielmehr müsste die Anzahl berechnet werden können, mit der man den Beobachtungen am nächsten kommen könnte. Es ist daher Sache des Durchführenden, sicherzustellen, dass die dimensionierten Unterverteilungen mit der Dynamik des Raums übereinstimmen. Dennoch besteht das Risiko, dass die Interpretierbarkeit dieser Klassifizierung verringert wird.

Fazit und Diskussion

Zusammenfassend gesagt zeigte das Projekt sowohl das Potenzial der Datenfusion zur Gewinnung neuer Informationen aus dem Zusammenspiel verschiedener Daten als auch die Grenzen eines solchen Verfahrens auf. Es ist wichtig, dass die gewählten Ausgangshypothesen zum Modellieren des Verkehrs tragfähig sind, das heißt mit den Voraussetzungen für die Datenfusion und der Logik der Bewegungen im untersuchten Bereich übereinstimmen.

Mit der Datenfusion können verschiedene Hürden überwunden werden, die sich beim Einsatz eines Verkehrsüberwachungssystems ergeben. Sie kann beispielsweise die Kosten begrenzen, die durch die Verwendung einer grossen Zahl von Sensoren entstehen würden. Gewissen Kombinationstypen muss man jedoch mit Vorsicht begegnen, beispielsweise, wenn die Fusion den Zugang zu Informationen ermöglicht, die absichtlich weggelassen wurden (mit denen z. B. eine Person identifiziert werden kann). Dieses Risiko ist gering, wenn man Flussschätzungen kombiniert, aber es ist vorstellbar, dass das System langfristig mit soziodemografischen Daten ergänzt werden könnte. Dann wäre es unabdingt notwendig, präventive Massnahmen zu ergreifen, damit die Privatsphäre der Nutzer nicht verletzt wird.

Ausserdem konzentrierten wir uns auf einen relativ kleinen Bereich. Um noch aussagekräftigere Ergebnisse zu erhalten, müsste diese Zone vergrössert werden. Dies würde aber auch bedeuten, dass mehr Akteure involviert und mehr Auflagen zu berücksichtigen wären. Eine genaue Festlegung der neuen Zone wäre deshalb wichtig, genauso wie nahezu vollständige Kenntnis der Interaktionen, die dort stattfinden. Die nachfolgenden Fragen

Remarques

L'hypothèse principale du modèle est que les variables suivent des lois de distribution connues. Cependant, la simplification en gaussienne est trop imprécise pour prédire correctement les débits. On pourrait choisir une loi de distribution plus adaptée à modéliser un trafic, telle que la loi de Poisson tronquée en zéro, qui a l'avantage de ne pas admettre de valeurs négatives.

Les bruits et erreurs de mesure ont été négligés dans la construction du modèle, mais il est clair à la vue des résultats qu'il serait difficile de remarquer la présence d'anomalies dans les données. L'idéal serait de ne pas imposer un nombre de distributions, comme cela a été fait ici où on a imposé la présence de 3 sous-distributions, mais au contraire, de calculer le nombre qui permettrait de s'approcher au plus des observations. Il conviendra ensuite à l'opérateur de déterminer si les sous-distributions dimensionnées sont cohérentes avec la dynamique de l'espace. Cependant, on court le risque de diminuer l'interprétabilité de cette classification.

Conclusion et discussion

Pour conclure, ce projet a montré à la fois le potentiel de la fusion de données pour l'obtention de nouvelles informations à partir de jeux de données distincts, et ses limites. Il est important de choisir de bonnes hypothèses de base pour modéliser le trafic, à la fois cohérentes avec les besoins de la fusion et la logique des déplacements dans le périmètre.

La fusion de données permet de remédier à de nombreux obstacles qui se posent lors de la mise en place d'un système de monitoring de trafic. Par exemple, elle peut limiter les coûts engendrés par la pose d'un grand nombre de capteurs. Il faut cependant prendre garde à certains types de combinaisons, par ex. lorsque la fusion permet de retrouver des informations qui avaient été omises volontairement (permettant par ex. d'identifier un individu). Ce risque reste faible lorsque l'on combine des estimations de flux, mais on peut imaginer à long terme vouloir enrichir le système de données sociodémographiques. Dans ce cas, il est essentiel de mettre en place un certain nombre de mesures préventives pour éviter toute intrusion dans la vie privée des usagers.

Enfin, nous nous sommes concentrés sur une petite zone, mais il est évident que pour inférer des caractéristiques plus pertinentes, il est nécessaire d'élargir le périmètre. Ceci s'accompagnera invariablement d'une multiplication d'acteurs et de contraintes. Il est donc essentiel d'assurer une définition précise de la nouvelle zone, avec une connaissance quasi-exhaustive des interactions qui la régissent. Les questions suivantes peuvent servir de base à une généralisation de l'architecture de récolte de données et des modèles de fusion:

- 1. Quels sont les usagers de la zone?** Il est essentiel de connaître la fréquentation de la zone en termes de modes de transport d'une part et de catégories sociodémographiques d'autre part. Les modes de transports utilisés permettent d'identifier les types de capteurs physiques

können die Grundlage einer Ausweitung des Datenerfassungssystems und der Fusionsmodelle bilden:

1. **Wer nutzt das Gebiet?** Es ist wichtig, dass bekannt ist, wer das Gebiet besucht – sowohl in Bezug auf die Verkehrsarten als auch auf die soziodemografischen Kategorien. Anhand der genutzten Verkehrsmodi kann bestimmt werden, welche physikalischen Sensoren installiert werden sollen, während soziodemografische Informationen nützlich sind, um eventuelle Verzerrungen des digitalen Systems zu berichtigen.
2. **Welche Transportinfrastruktur gilt es zu berücksichtigen?** Die Nutzung der Infrastruktur hat einen direkten Einfluss auf die Dimensionierung des Beobachtungssystems (Art, Anzahl und Position der Sensoren).
3. **Was sind die möglichen Routen und Zugänge?** Kenntnis der Infrastruktur reicht nicht aus, um sich ein vollständiges Bild der Mobilität zu machen, insbesondere, was Fußgänger betrifft, da diese sich freier bewegen können. Die wichtigsten Ein- und Ausgänge des Gebiets sowie Alternativrouten müssen unbedingt identifiziert werden.
4. **Was sind die Anziehungspunkte des Gebiets?** Unter Anziehungspunkten versteht man Einrichtungen, Geschäfte und Infrastrukturelemente, die über alle Verkehrsmodi hinweg Nutzer anziehen und zahlreiche Bewegungen erzeugen. Durch das Untersuchen dieser Punkte können die optimalen Standorte für Sensoren identifiziert werden, sodass ein möglichst grosser Anteil des Verkehrs erfasst werden kann.
5. **Wo bestehen Lücken?** Ein lückenloses Wissen über alle Aspekte des Gebiets ist nicht realistisch, und auch die leistungsfähigsten Sensornetze können nicht den gesamten Verkehr fehlerfrei darstellen. Es ist jedoch möglich, komplexe Aspekte, die bei der Erkennung der Verkehrsflüsse systematische Fehler generieren könnten (z. B. Hindernisse im Erfassungsbereich), vorauszusehen und dann Massnahmen zu ergreifen, um diese vor Projektbeginn zu beheben.

Diese Liste deckt selbstverständlich nicht alle Fragen ab, aber sie erlaubt es, ein System zur optimalen Datenerfassung für grosse und dynamische Gebiete aufzubauen. Zudem hilft diese einfache vorgängige Analyse dabei, potenzielle Problembereiche zu identifizieren. Dadurch wird eine gewisse Aussagekraft der Daten sichergestellt, und nötigenfalls können Massnahmen zur Verbesserung ihrer Qualität getroffen werden.

qui doivent être mis en place, tandis que la connaissance des catégories sociodémographiques est utile pour corriger d'éventuels biais du système digital.

2. **Quelles sont les infrastructures de transport à prendre en compte?** La fréquentation des infrastructures aura un impact direct sur le dimensionnement de l'Observatoire (type, nombre et position des capteurs).
3. **Quelles sont les itinéraires et accès possibles?** La connaissance des infrastructures n'est pas suffisante pour avoir une vue complète de la mobilité, notamment en présence de piétons qui sont peu contraints par l'espace physique. L'identification des entrées et sorties principales dans la zone, ainsi que des alternatives d'itinéraires, est essentielle.
4. **Quels sont les pôles d'attractivité de la zone?** Par pôles d'attractivité, on entend tous les équipements, commerces et infrastructure qui attirent et génèrent de nombreux déplacements, tous modes confondus. L'étude de ces pôles permet d'identifier naturellement les localisations optimales des compteurs, où un maximum du trafic peut être détecté.
5. **Quelles sont les lacunes?** Il est évident qu'une connaissance exhaustive de tous les aspects de la zone est impossible, et même le plus performant des réseaux de capteurs ne peut représenter l'intégralité du trafic sans erreur. Cependant, il est possible d'anticiper et identifier des points complexes qui risquent de générer des erreurs systématiques dans la détection des flux (e.g. obstructions dans le champ de détection), et implémenter des mesures correctives en amont.

Cette liste de questions n'est bien entendu pas exhaustive, mais permet de mettre en place une architecture de récolte de données optimale pour des périmètres larges et dynamiques, et éviter des écueils qui auraient pu être identifiés avec une simple analyse préliminaire. On garantit ainsi une certaine pertinence des données, et au besoin des mesures de correction efficaces pour en améliorer la qualité.

Références

- [1] R. Rollier, «A smart data product to monitor traffic», 2016. [Online]. Available: <https://ict.swisscom.ch/2016/12/a-smart-data-product-to-monitor-traffic/>. [Accessed: 11-Jan-2019].
- [2] Lang Anthony, «Développement d'un observatoire de la mobilité utilisant des données de téléphones mobiles et d'autres senseurs: cas d'étude à Pully», EPFL, 2017.
- [3] F. C. Chan, «Using Mixture Models for Clustering», 2015. [Online]. Available: <http://tinyheero.github.io/2015/10/13/mixture-model.html>. [Accessed: 06-Jan-2019].
- [4] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild, «Bayesian correlated clustering to integrate multiple datasets», Bioinformatics, vol. 28, no. 24, pp. 3290–3297, Dec. 2012.
- [5] A. P. Dempster, N. M. Laird, and D. B. Rubin, «Maximum likelihood from incomplete data via the EM algorithm», J. R. Stat. Soc. Ser. B, pp. 1–38, 1977.
- [6] F. C. Chan, «Fitting a Mixture Model Using the Expectation-Maximization Algorithm in R», 2016. [Online]. Available: <http://tinyheero.github.io/2016/01/03/gmm-em.html#fitting-a-gmm-using-expectation-maximization>. [Accessed: 06-Jan-2019].