Whitepaper
Mobility Pattern Recognition (MPR) und Anonymisierung von Mobilfunkdaten

Andreas Neumann[1] und Michael Balmer[2]

[1] Senozon Deutschland GmbH, Berlin, Deutschland, andreas.neumann@senozon.com
[2] Senozon AG, Zürich, Schweiz, michael.balmer@senozon.com

Einleitung

Trotz des immer mehr im digitalen Raum ablaufenden Wirtschaftslebens, ist die Bevölkerung heute physisch so viel unterwegs, wie nie zuvor. Der Pendelverkehr und mit ihm die Verkehrsströme haben ein noch nie dagewesenes Volumen erreicht. Verkehrsstromanalysen liefern für verschiedene Wirtschaftszweige und auch für Bereiche der öffentlichen Versorgung und Verwaltung Einsichten von unschätzbarem Wert. Vier Wirtschaftsbereiche, die besonders auf diese Analysen angewiesen sind, möchten wir im Sinne einer beispielhaften Aufzählung herausstreichen, um den generierten wirtschaftlichen Mehrwert zu verdeutlichen.

- Der Handel, der bei vielen Dienstleistungen und Konsumgütern auch heute noch – entgegen der allgemeinen Digitalisierungstendenz – in der physischen Welt stattfindet, ist darauf angewiesen, die Wirtschaftlichkeit von Standorten schon vor der Eröffnung eines neuen Lokals abzuklären. Dies erklärt sich von selbst, wenn man die bedeutenden Investitionen bedenkt, die in Verkaufslokale getätigt werden für Innen-ausbau, Personalrekrutierung und -ausbildung, Beschilderung und andere werbewirksame Massnahmen zum Bekanntmachen des neuen Standorts und so weiter. Typische Fragen solcher Kunden sind: Wo eröffne ich am besten mein Take-Away Lokal? Welche Öffnungszeiten machen Sinn? Wo bietet ein Bankomat am meisten Mehrwert? Wie gut ist der Standort per Auto erreichbar unter Berücksichtigung der regelmässigen Verkehrsbelastungen? Wo wünschen sich unsere Kunden Support-Standorte?
- Für die Werbewirtschaft ist es hochrelevant, ihre Werbeflächen mit fundierten Angaben zum erwarteten Publikum anzubieten. Die Erwartungshaltung der werbenden Unternehmen ist angesichts der im digitalen Raum zielgenauen Ansprache einzelner Kunden stark gestiegen. Wer Plakatwände bucht, möchte heute wie im digitalen Raum sehr detailliert und möglichst merkmalsreich steuern und wissen, wen er mit seiner Werbenachricht erreicht. Typische Fragen solcher Kunden sind: Welche Bevölkerungsgruppen kann ich mit dieser Fläche ansprechen? Finde ich zu bestimmten Tageszeiten ein ganz anderes Publikum vor (sehr relevant für minutengenau steuerbare digitale Werbeplakate)? Welche Bedürfnisse hat das Publikum an diesem Standort am ehesten? Oder von den Werbenden selbst: Wo ist mein Werbebudget am gewinnbringendsten investiert? Wo wohnt mein typischer Kunde und welchen Pendelweg legt er zu welchem Arbeitsort zurück?
- Die Immobilienbranche ist gleich in zweierlei Hinsicht auf solche Auswertungen an-gewiesen. Einerseits ermöglichen sie ihnen, ihre bestehenden Geschäftslokale mit qualifizierten Anpreisungen besser zu vermarkten. Anderseits können Neubauprojekte bzw. grosse Renovationen bereits im Voraus auf ihre Wirtschaftlichkeit geprüft werden.

- Auch die öffentliche Hand bzw. der öffentliche Verkehr profitiert stark. Neben Analysen zur aktuellen Lage, wie zum Beispiel der aktuellen Lärmbelastung verschiedener Standorte, sind auch Auswertungen zu Zukunftsszenarien gefragt, vor allem beim öffentlichen Verkehr und der Städteentwicklung. Typische Fragen solcher Kunden sind: Welchen Einfluss hat die Erschliessung eines bestimmten Areals auf das umliegende Verkehrsaufkommen? Wie wirken sich welche Gegenmassnahmen aus? Wo würde eine neue ÖV-Linie auch genutzt und würde sie den Individualverkehr reduzieren? Wie hoch ist die Lärmbelastung in bestimmten Gebieten und wie verändert sich diese Belastung mutmasslich mit verschiedenen Projektvarianten?
- Nicht zuletzt – und wichtiger denn je – sind vertiefte Kenntnisse zum typischen Mobilitätsverhalten einer Bevölkerung elementar wichtig, um effizient und wirksam Massnahmen zu entwickeln, um Ausbreitungen von Krankheiten und Entwicklungen von Epidemien und Pandemien zu verhindern oder verlangsamen (Stichwort: «Flatten the Curve»).

Hier bieten Mobilfunkdaten neue Möglichkeiten, diese Fragen auf fundierten und kontinuierlich erhobenen Grundlagen zu beantworten. Am einfachsten wäre es, das Verhalten der Bevölkerung anhand realer Personen zu modellieren, indem man zum Beispiel die Kunden eines Telekommunikationsunternehmens unter Beibehaltung ihrer sozio-demographischen Merkmale und Bewegungsmuster als Datenbasis verwendet und lediglich die direkt identifizierenden Merkmale (Name, Adresse, Geburtsdatum etc.) verwirft., d.h. löscht. Eine solche grossflächige Bearbeitung von merkmalsreichen – wenn auch anonymisierten – Sachdaten birgt aber ein inhärentes Risiko der Re-Individualisierung.

In der Modellierung von dynamischen Systemen – im speziellen in der Modellierung von Verkehrsnachfragen – wird seit etwa 2 Jahrzenten das Konzept der «synthetischen Bevölkerung» angewendet. Hierbei werden auf der Basis künstlich generierter Personen, die statistisch betrachtet die reale Bevölkerung in ihrer Demographie und Soziodemographie wiederspiegeln, deren Mobilitätsbedürfnis modelliert. Das heisst, sie bewegen sich wie die reale Bevölkerung durch den Raum, gehen morgens zur Arbeit, machen Freizeit oder gehen vor Ladenschluss noch einkaufen. Sie nutzen die zur Verfügung stehenden Verkehrsangebote, fahren mit dem Auto oder Fahrrad und suchen sich die passendste Route, um ihre Ziele zu erreichen. Das heisst, sie repräsentieren die Mobilität der Bevölkerung, ohne dass hierbei reale Personen verfolgt werden.

Diese Modelle basieren typischerweise auf aufwändig durchgeführten Befragungen, die in ihrem Umfang und somit auch in ihrer Modellgüte entsprechend limitiert sind. Was aber, wenn diese Methode auch auf «Quasi-Befragungen» durch Verarbeitung von Mobilfunkdaten so kombinierbar sind, dass einerseits der Umfang der «Befragten» massiv steigt und hierbei die Methodik der Anonymisierung durch Verwendung einer synthetischen Bevölkerung weiterhin bestehen bleibt? Im Folgenden wird hier diese kombinierte Methodik vorgestellt.

Mobility Pattern Recognition: Interpretation der realen Personen- und Bewegungsdaten

Das Modell ist für aussagekräftige Analyseergebnisse darauf angewiesen, dass der Simulation möglichst lebensnahe Daten zu Grunde liegen. Als Basis dienen ihr zum einen öffentlich zugängliche Daten wie etwa Bevölkerungsstatistiken, Fahrpläne, Gebäude- und Nutzungsdaten, sowie die durch die Mobilfunkgeräte erzeugten «Signalisierungs-Events».

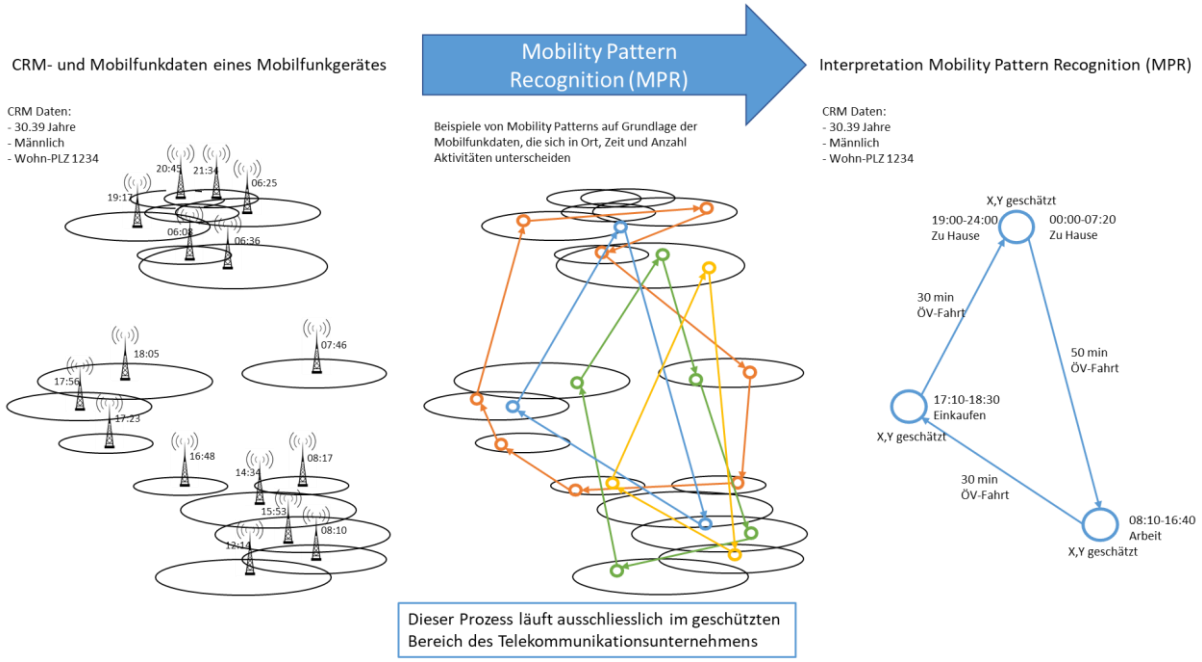

Abbildung 1: Beispielhafte Signalisierungsdaten eines Mobilfunkgerätes mit CRM-Daten und der daraus abgeleitete MPR-Bewegungspfad

Zunächst wird auf den gesicherten Systemen des Telekommunikationsanbieters aus den vorhandenen realen Event-Logs (Aktivität eines Mobilfunkgeräts, zum Teil mit CRM Daten angereichert, mit Zeitstempel und Angaben zum Sendemast und/oder Transmitter) mittels des durch die Senozon entwickelten MPR-Algorithmus («Mobility Pattern Recognition») ein wahrscheinlicher Bewegungspfad abgeleitet (blauer Pfad in Abbildung 1). Die realen Event-Logs erlauben verschiedene Interpretationen. Andere mögliche, aber weniger wahrscheinliche Pfade (oranger, grüner und gelber Pfad in der Grafik) werden verworfen.

Zudem werden aus den wahrscheinlichen Bewegungspfaden basierend auf der zeitlichen Verteilung der Aufenthalte über den Tag wahrscheinliche Aktivitäten abgeleitet (z.B. zuhause, Arbeit, Weg, Einkaufen etc.). Wichtig ist, dass die wahrscheinlichen Aktivitäten also nicht an-hand der tatsächlichen Aktivitäten ermittelt werden, sondern aufgrund von statistischen Wahrscheinlichkeiten.

Die Methode wendet hier drei Anonymisierungsschritte an:

1. die Generalisierung von Datenfeldern (z.B. Alter 30-39 statt Alter 34),
2. Unschärfen bei der Erfassung der Events (grosses Sendegebiet eines Sendemastes, keine Triangulation) und
3. Auswahl von Wegen (ohne Routen) und Aktivitäten auf Basis von statistischen Schätzungen anstelle von tatsächlichen Gegebenheiten.

Im Ergebnis resultiert ein zwar aussagekräftiges Bewegungsprofil mit Aktivitätenketten, das den Gesetzmässigkeiten der realen Bevölkerung folgt, das aber so gut wie gar nie den tatsächlichen Gegebenheiten entspricht. Vor dem Export werden die Datensätze mit den Bewegungsprofilen und die Aktivitätenketten zudem noch weiter anonymisiert (siehe folgende Abschnitte).

Start-Ziel Matrizen mit Verkehrsmittel und standardisierten Zeitfenstern

Aus dem Bewegungspfad werden sodann anonymisierte Start-Ziel Matrizen abgeleitet (siehe auch Abbildung 2). Hier-zu wird der Bewegungspfad übersetzt in einen einzigen binären Eintrag (0 oder 1) in einer geographischen Start-Ziel Matrix, wobei pro Stunde und Verkehrsmittel jeweils eine eigene solche Matrix angelegt wird. Die allenfalls vorhandenen CRM-Daten werden bei der Erstellung dieser Matrizen vollständig verworfen, das heisst die Matrizen geben keine Auskunft über die Reisenden selbst, wie etwa Alter und Geschlecht. Zudem ist jeder Weg wie erwähnt nur binär gezählt, d.h. es wird festgehalten, ob eine Reise von geographischer Start- zu Ziel-zelle stattfindet, aber nicht etwa wie viele Personen eine solche Reise zurücklegen. Die Zuweisung zu einem bestimmten Verkehrsmittel erfolgt mittels einer Schätzung eines wahrscheinlichen Verkehrsmittels aufgrund des gegebenen Verkehrsangebots. Das von den realen Individuen tatsächlich gewählte Verkehrsmittel wird dabei nicht berücksichtigt.

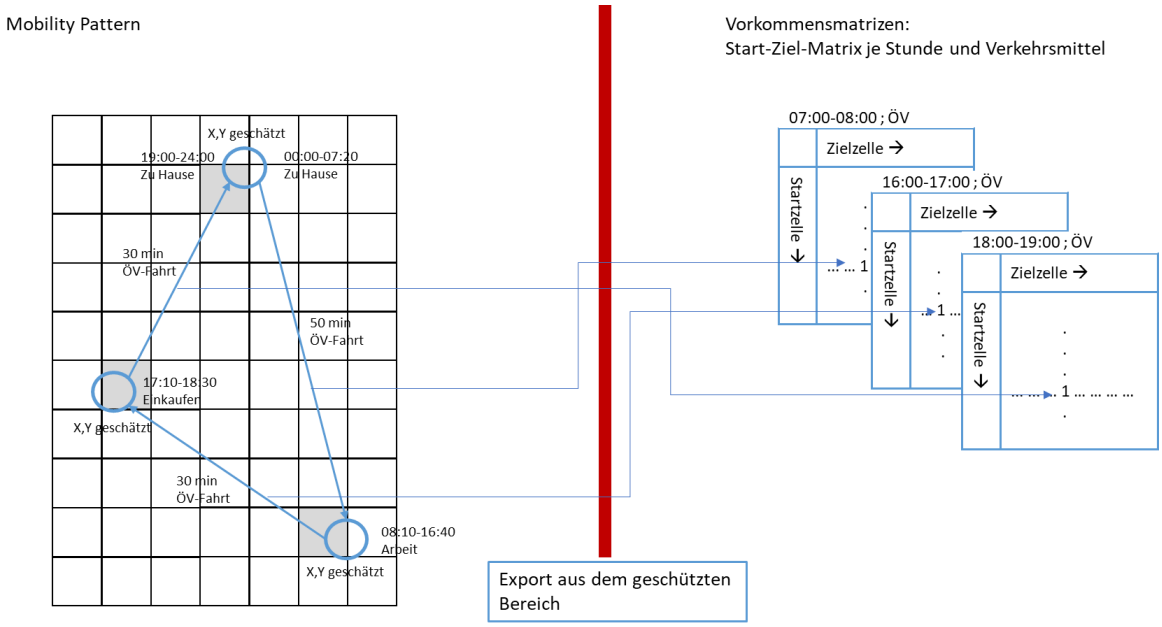

Abbildung 2: Transformierung eines MPR-Bewegungspfades zur Einordnung in eine Matrix

Hierdurch resultieren Matrizen, die zwar noch eine Aussage darüber erlauben, wann ungefähr zwischen welchen geographischen Punkten überhaupt Reisen stattfinden und mit welchen (wahrscheinlichen) Verkehrsmitteln. Sie enthalten aber keinerlei Angaben mehr dazu, wie viele Personen diese Reise unternommen haben und welche soziodemographischen Merkmale die Reisenden aufweisen.

Synthetische Individuen mit anonymen Aktivitätenketten

Parallel zum oben aufgeführten Prozessschritt werden die Aktivitätenketten isoliert und weiterbearbeitet. Die Ortsangaben der einzelnen Aufenthalte werden dabei gelöscht, beibehalten werden die Anzahl und Dauer der detektierten Aufenthalte, ob es sich um einen wahrscheinlichen Übernachtungsort handelt und die geschätzten Distanzen zwischen den Aktivitäten. Diese isolierten Aktivitätenketten werden zunächst in einem Datenpool gespeichert, wobei die realen CRM-Daten in ihrer generalisierten Form (z.B. Alter 30-39 statt 34) vorerst noch erhalten bleiben (siehe Abbildung 3).

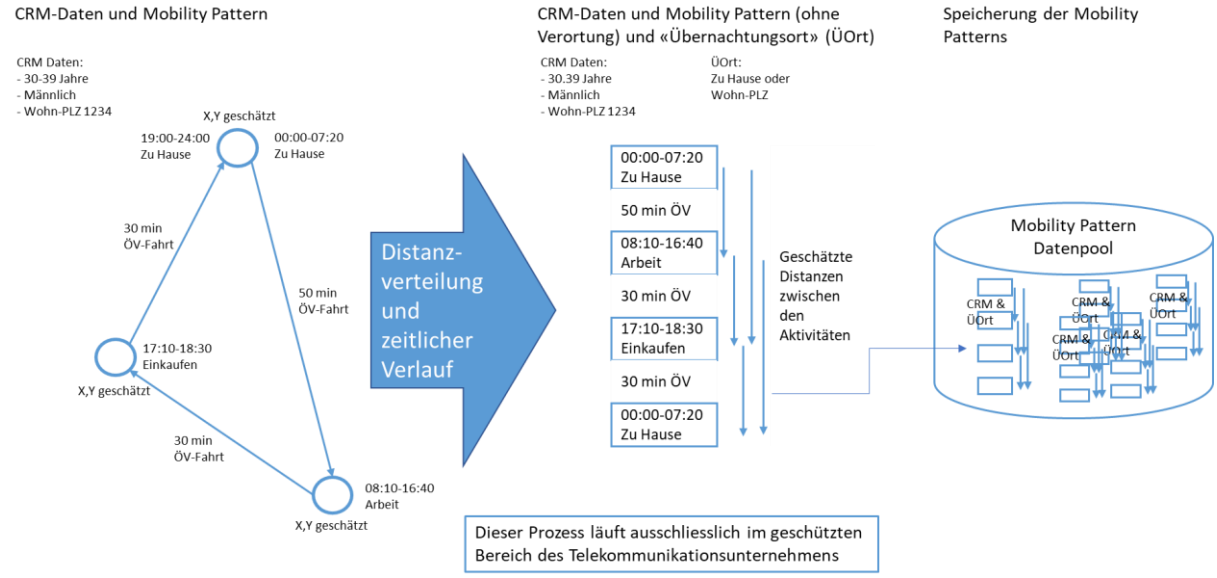

Abbildung 3: Transformierung eines Mobility Pattern Recognition-Bewegungspfades in eine Aktivitätenkette

In einem nächsten Schritt werden aus der von Senozon bereits zuvor auf Basis anderer Daten simulierten, synthetischen Bevölkerung einzelne simulierte Individuen ausgewählt. Für diese Individuen werden aus dem Datenpool mindestens 30 verschiedene, aufgrund der CRM-Daten möglichst «nahe» Aktivitätenketten ausgewählt. Aus diesen 30 Ketten wird nach Zufallsprinzip eine Kette gezogen und dem simulierten Individuum als authentisches Verhalten zugewiesen (siehe Abbildung 4).

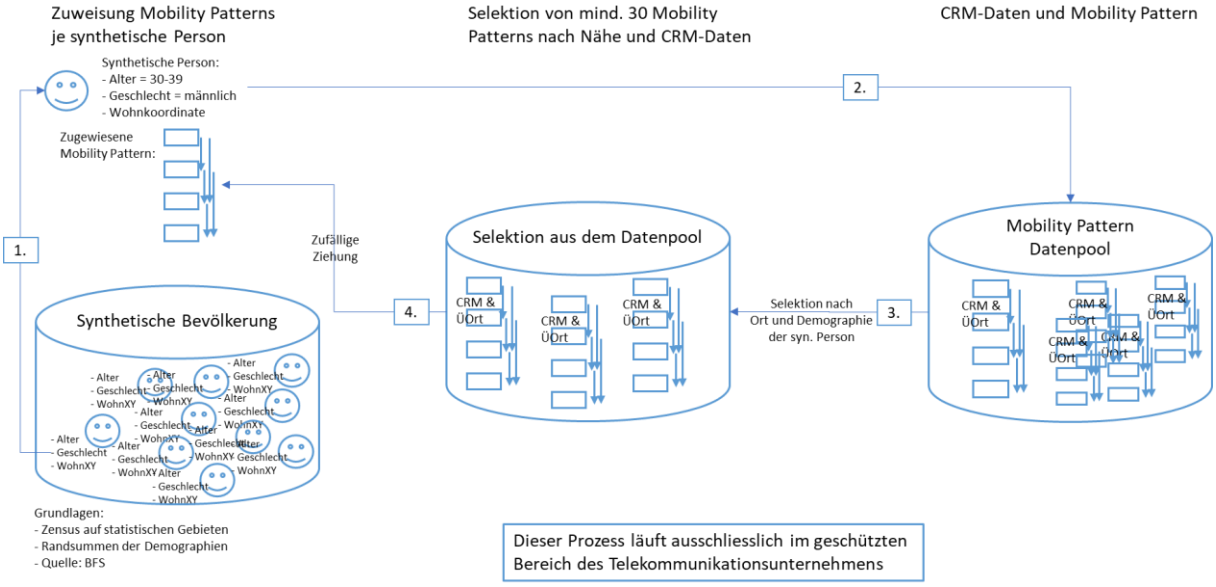

Abbildung 4: Zulosung der Aktivitätenketten zur synthetischen Bevölkerung

Die Anonymisierung geschieht bezüglich dieser Daten mittels Ersetzung der generalisierten CRM-Daten durch zufällig zugeloste fiktive «Personendaten» der synthetischen Individuen. Die Aktivitätenketten selbst werden vor der Übermittlung an Senozon zwar nicht weiterbearbeitet, sie entsprechen aber wie erwähnt so gut wie gar nie den tatsächlichen Gegebenheiten. Dies ergibt sich aus dem Prozess ihrer Erstellung:

- Zunächst einmal ist es aufgrund der Unschärfe der Ausgangsdaten und ihrer Verarbeitung (grosses Sendemastgebiet, Verwerfung anderer möglicher Bewegungspfade, grosse Altersgruppe) unwahrscheinlich, dass die Interpretation die Realität akkurat abbildet.
- Dann ist es unwahrscheinlich, dass das synthetische Individuum in seinen individuellen Merkmalen genau der zur Aktivitätenkette zugehörigen Person entspricht, denn die Aktivitätenketten werden nach dem Zufallsprinzip aus einer Gruppe von 30 bezüglich CRM-Daten möglichst «nahen» Aktivitätenketten gezogen.

Durch diese Unschärfen und Zufallsziehungen sind die Aktivitätenketten nicht mehr aussagekräftig genug, um eine Zuordnung zu ermöglichen.

Daten-Export

Die Start-Ziel Matrizen sowie die synthetischen Individuen mit ihren Aktivitätenketten werden schliesslich durch den Telekommunikationsanbieter zur Weiterverarbeitung für die entsprechenden Mobilitätsmodelle exportiert. Die Start-Ziel Matrizen dienen dabei zur Verbesserung der im Model eingepflegten Bewegungslogik der synthetischen Individuen. Die Aktivitätenketten dienen dazu, dass die Verkehrsströme der synthetischen Bevölkerung bezüglich ihrer soziodemographischen Merkmale möglichst authentisch sind, d.h. möglichst analog der tatsächlichen Bevölkerung.