CIHT Dubai Online Seminar – Bulletin

Luis (Pilo) Willumsen, Managing Partner Nommon Solutions and Technology,

About the Event

New data sources and collection techniques have become available enabling larger sample sizes and increased temporal coverage. The techniques acquire data using a variety of methods including sensors for Bluetooth and WiFi signals, public transport smart cards, smartphone apps, mobile network data and vehicle telematics. These are passive, non-disruptive, forms of data collection that take advantage of the digital traces generated by a range of devices.

Mobile phone network data (MND), the digital traces our phones leave behind, can be processed to trace movements, and generate anonymised origin destination matrices and other mobility indicators, 24 hours a day 365 days in a year.

MND has strengths and limitations and these will be discussed in the Seminar. New techniques based on machine learning have removed most limitations and their use in transport models is now common practice in countries such as the UK, the US and Spain.
The seminar will discuss these issues and provide a 'state of the art" review of techniques and applications, including large and small transport models in different countries.

About the Speaker

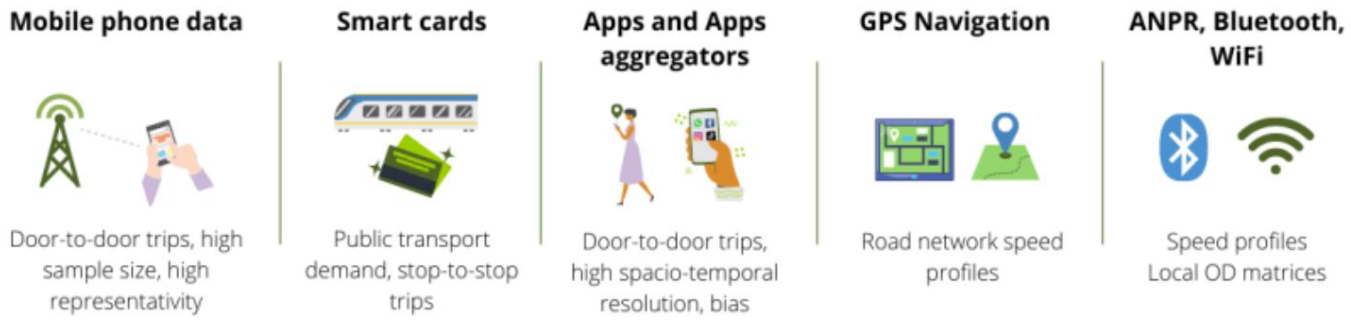
Luis (Pilo) Willumsen is an internationally recognised expert in Transport and Traffic modelling with over 35 years of experience as a consultant, transport planner and researcher with a valued academic career. He is co-author of "Modelling Transport" published by Wiley; its fifth edition will appear later this year. He also published "Better Traffic and Revenue Forecasting", a book dealing with the critical task of delivering demand and revenue projections for transport concessions. He is Managing Partner of Nommon Solutions and Technology, a company that uses big data sources to develop mobility insights. In that role he has pioneered the use of big data, in particular from Mobile Phone Networks, to deliver useful and innovative transport models.

Seminar

In his welcoming remarks, **Martin Tillman**, Chair CIHT Dubai, welcomed attendees and the speaker of the event.

The webinar commenced with **Luis (Pilo) Willumsen** providing the overview of the New data sources.

Luis (Pilo) Willumsen explained that Mobile phone network data (MND), the digital traces left behind by our phones, can be processed to trace movements, and generate anonymised origin destination matrices and other mobility indicators, 24 hours a day 365 days in a year. Cities are currently equipped with numerous sensors. Specifically, video cameras which employ automatic number plate recognition, and there are also Bluetooth and Wi-Fi sensors present. These data from hey have sensors have been used at various studies primarily for monitoring the movement of people within buildings. Models have been developed in at least four countries, including the United States, Spain, United Kingdom and Brazil using real time data. In UAE, Abu Dhabi, uses it to supplement data and support strategic transport model development.



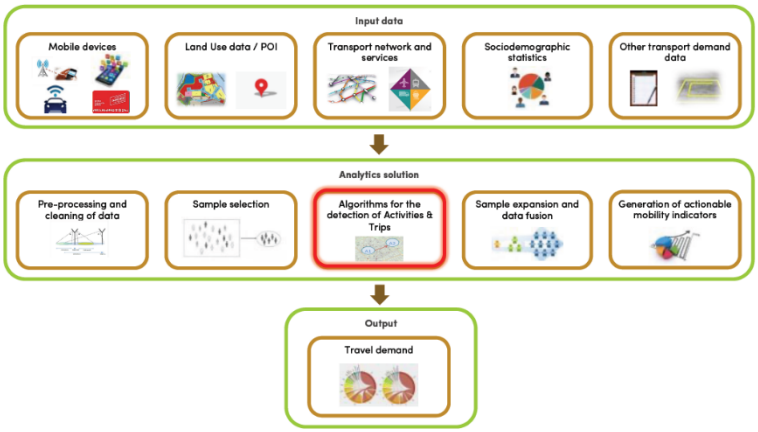
The process begins by obtaining raw data from the mobile phone operator, identifying locations where the mobile phone has stayed for a significant duration. Based on the duration of stay, it is inferred that these locations may likely be either the place of work or residence. Analysis involves recognising patterns of travel between places, including any gaps observed. We then analyse the land use of these locations, categorising them as residential areas, workspaces, or other types. Analysis infer the nature of the activity by considering the land use of each location and assessing the frequency and duration of stays at each stop and reconstruct the entire trip.
Significant advantage of mobile phone data is the availability of information for every day of the year. This approach offers the benefit of tracking a phone's movement over the course of a month, providing a high level of confidence in determining the residence and workplace locations due to the regularity of these trips at consistent times.

The size of the sample obtained from mobile phone data is considerably larger than that obtained through conventional roadside interviews or household travel surveys. Despite the inherent noise and errors in mobile phone data, techniques have been developed to clean and scrub the data and apply techniques like data fusion integrating various data sources.
Mobile phone data also has limitation and does not fully replace traditional surveys. Notably, it lacks information on vehicle ownership, driving licenses, gender, and income. Although income can be inferred from residential areas, it is not directly obtained. The limitations extend to the identification of only a few activities and trip purposes, and the data does not offer the depth of information achievable through state preference surveys, such as willingness to pay. Limitations also exist in in distinguishing between closely situated roads, such as parallel roads separated by only 50 or 100 meters. In urban areas, the speeds of different types of vehicles are often too similar to discern. Additionally, very short trips, particularly walking trips that occur within a single mobile phone cell, may go unnoticed as they do not involve movement.

Overcoming the limitations of MND
To address these limitations, **Luis (Pilo) Willumsen** explained the need for estimates of these short trips, which are typically intrazonal trips in modelling terms. These trips can be significant, especially when promoting active modes of transportation. Overcoming these constraints involves the use of data fusion, as previously described, and the application of machine learning techniques.

Big Data for Transport Models

10th October 2023



The **Luis (Pilo) Willumsen** discussed how their work was later incorporated into modelling software packages such as Saturn, focusing on estimating or updating origin-destination trip matrices using traffic counts. The objective is to estimate the trip matrix (TJ), which may have a prior historical or synthetically generated matrix applying entropy maximisation approach to achieve this, aiming to maximise a specific function while adhering to constraints. Each constraint, corresponding to a traffic count, ensures that the volume on a link equals the trips passing through it.

Using Machine Learning

Machine learning is employed to enhance data quality, enrich datasets, and build forecasting models. While we have not yet delved into generative artificial intelligence, there is a focus on supervised machine learning for classification and regression problems. This involves testing approximately eight different machine learning models, including decision trees, random forest, and gradient boosting. To address the challenge of the "black box" nature of machine learning models, efforts are made to enhance interpretability.

The general methodology for machine learning involves a training phase, analogous to a calibration phase in transport models. For data enrichment, the speaker explains the use of household travel surveys to train machine learning models to estimate characteristics such as gender, age, and income based on trip information. The training set typically uses 80% of the data, leaving 20% for testing.

In the application phase, the trained machine learning model, or the best-performing one, is applied to mobile phone data, predicting gender, age, trip purpose, income, etc., based on detected trip characteristics. A validation exercise is conducted to compare the model's results with the original survey data, ensuring alignment.

Use Case : Mallorca Airport access

Luis (Pilo) Willumsen provided an example of applying this methodology to Mallorca Airport, where machine learning models were trained on mobile phone data to identify and classify travellers, their modes of access to the airport, and estimate the length of their stay. The models proved effective, particularly in compensating for the skewed data and subsequently being able to represent the majority of tourists with some certainty.

Use Case: Elinsky, Finland Bike Sharing Scheme

The use of machine learning has been extended to forecasting and optimising new transportation modes, such as bike-sharing systems. In the case of Elinsky, Finland, mobile phone data, bike operator information, and weather data were integrated to assess the impact of expanding the bike-sharing system without detracting from public transport demand.

Use Case : Short-term demand estimation for rail

Another application involved collaboration with a train manufacturer and operator, Simons, to optimise train operations. Machine learning models were trained on past operational data, including ticketing and weight data, to forecast demand for the next 90 minutes and 24 hours ahead. The objective was to balance service quality, punctuality, and energy efficiency by suggesting adjustments like inserting or removing trains, modifying dwell times, and incorporating

intermediate turnarounds.

The results of these applications indicated the effectiveness of machine learning in providing short-term operational suggestions, reducing costs, and optimising service levels without compromising passenger satisfaction. The summary emphasises the importance of using machine learning in specific applications, especially for new transportation modes, where traditional models may fall short in capturing complex factors influencing user behaviour.

In conclusion, the **Luis (Pilo) Willumsen** emphasises the significant role of big data, particularly from mobile phones and smart devices, in revolutionising transportation models. The availability of vast and current datasets allows for improved model accuracy, rapid updates, and the detection of evolving trends and behavioural changes. This is exemplified by the application of mobile phone data during the pandemic, where the World Bank sought insights into the impact on public transportation usage in Latin American cities, specifically Bonosciis and Bogota.

The speaker emphasises the importance of processing and validating mobile network data, suggesting that these steps may be more crucial than the raw data quality itself. Utilising machine learning, the cleaning and validation process becomes more efficient, enabling the identification of trends and the formulation of strategies to address or adapt to changing circumstances.

The talk also highlights the versatility of machine learning in enriching datasets, as demonstrated in the World Bank project where it was used to identify transportation modes in large cities. The technology proves especially valuable when dealing with new challenges posed by emerging mobility technologies, where traditional models may fall short.

Questions

During data cleaning, if trips not ending in the study area are within the day are removed, are we excluding long distance trips and trips to airports, high speed stations?

The speaker addressed concerns raised regarding the potential bias in data due to the exclusion of long-distance trips, such as those to airports or high-speed stations. Broader issue encountered in Brazil, where the conventional application of the technique involved producing daily origin-destination trip matrices that might not capture longer trips lasting multiple days. The speaker acknowledged the challenge, especially for countries like Brazil with extensive coverage by a single mobile operator, resulting in very long trips. To address this, the speaker highlighted the modification of algorithms to accommodate longer trips, a necessity recognised by consultants and clients in Brazil. The speaker also explained how this adjustment is not as relevant in the UK and Spain unless the study extends beyond their borders, as seen in international studies involving flights where mobile phone data helps identify the country of destination. Overall, the speaker emphasised the adaptability of algorithms to different contexts and the importance of capturing comprehensive travel patterns.

What is the minimum time resolution of data? Is it hourly based?

The speaker discussed the variability in data availability across countries and operators, highlighting that in Spain, data is obtained every 3 to 5 minutes, while Brazil is progressing toward more frequent data collection. When addressing the question about the minimum time resolution of data, the speaker suggested a one-hour interval as the minimum for producing trustworthy trip matrices, emphasising the challenge of trusting shorter time periods due to potential data inaccuracies.

Regarding the issue of data storage, the speaker explained that while traditional models do not impose additional storage requirements, the processing of mobile phone data does demand more storage capacity. The data processing is often performed within the mobile phone operator's servers to address privacy concerns, with restrictions on the release of individual information. The speaker clarified that the storage capacity required is substantial, but it is typically managed by the mobile phone operators, either on-ground servers or cloud-based servers like Amazon Web Services and Google, tailored for specific operators. Overall, the speaker downplayed concerns about excessive storage demands, especially when compared to the requirements of large language models, which demand significant space for their billions of parameters.

Did the Siemens Trains have automated passenger counts at each door?

The speaker clarified that the public transport system discussed did not have automated passenger counts at each door. Instead, the system relied on weighing the vehicles to determine the load of different carriages. The speaker mentioned that, in the UK, screens display information about the load on each carriage without the need for direct passenger counting.

Regarding recommendations for public transport operators, the speaker expressed uncertainty about specific measures but highlighted the value of ticketing information and smart card data for their work. The speaker suggested that incorporating passenger counters, especially for exiting passengers, could enhance the accuracy of their origin-destination matrices. However, the decision to introduce such measures would depend on the operational needs and the perceived value for the transport operators. The speaker emphasised their current practice of utilising existing data collected for operational purposes, without requiring additional data specifically for their research.

Would it be possible to share mode parameters used to identify the modes through ML?

The speaker emphasised that machine learning models do not generate bold parameters to share. While machine learning models have hyperparameters for calibration and refinement, they operate with a network of weights. In the case of random forests, these weights move through the model, but the focus is on the usefulness of the model itself, not the specific parameters. The speaker drew a parallel to models like CHATGPT, stating that knowing the billions of parameters is not useful; what matters is the effectiveness of the machine learning model, making it distinct from logic or gravity models where parameters might be directly handled.

What do you see is the required change of skills for a future transport modeller, given the emergence and growth in Big Data as an input? Is it about data science and ML? Is it about technical traditional skills? Is it about more general skills such as engagement with end users? And don't say it's about ALL!!

The speaker highlighted the evolving skill set required for future transport modelers, attributing it to the emergence and growth of big data as an input. They emphasised the need for a diverse skill set, encompassing data science, machine learning, traditional technical skills, and general skills such as user engagement. The comparison was drawn to the expanding complexity in everyday life, where constant learning is essential. The speaker acknowledged that the field now requires understanding machine learning, which, while new, can be specialised, much like other aspects of transport modelling. Overall, the message conveyed was that a comprehensive skill set is necessary for contemporary transport modelers