Human Behaviour Analysis Through Probabilistic Modelling of GPS data

In urban city planning, it is becoming increasingly difficult to improve and maintain the inhabitants’ quality of life and security due to the rising number of the population. Human behavioural characteristics and movement understanding can be an important tool to help assure improvement in the realm of urban planning. However, particularly when examining automated geolocated data (for example GPS data), studies in the field of human movement analysis are uncommon and often require labels that are difficult to find in real world applications. This is possibly due to the challenges associated with mining and analysing GPS data. It is dynamic, highly-irregularly sampled, noisy, and its data collection can be frequently interrupted due to environmental factors causing missing data. Moreover, different time series trajectories representing different journeys often vary with the number of observations adding challenges to the problem of comparing, clustering and grouping the data sequences. Nevertheless, the lower sensing modality of GPS data has less ethical and privacy concerns to other behaviour monitoring sensors such as CCTV cameras, yet can hold very rich information on behavioural characteristics making it an attractive data source to study.

In order to understand the data, compare between multiple trajectories, and bypass the previously mentioned challenges we opted to design a generative probabilistic model for GPS data summarising the high dimensional time series information with the models parameters and states. Using only the trajectory sequence \( X = (x_1, x_2, \ldots, x_T) \) where \( x_t \in \mathbb{R}^3 = (time, longitude, latitude) \), the feature sequence \( Y \) is calculated to describe the trajectory path where \( Y = (y_1, y_2, \ldots, y_T) \), \( y_t \in \mathbb{R}^4 = (longitude, latitude, h, v) \), \( h \) is the hour of day and \( v \) is the velocity calculated using the Haversine equation. Human movement patterns are complex making them difficult to describe with a single set of parameters. Therefore, the generative model structure may better resemble the mechanics of a switching model with parameters optimised for each state in the data sequence corresponding to different behaviour patterns. We propose a novel Non-Homogeneous Vector Autoregressive infinite Hidden Markov (NH-VAR-iHMM) model. Building on the theory of the VAR-HMM structure, we introduce a new discrete and independent semi-markov variable \( \tau = (\tau_1, \tau_2, \ldots, \tau_T) \) where \( \tau_t = i \). This acts as a parent to the state indicator variable \( z \) in the Probabilistic Graphical Model (PGM) to represent environmental factors that influence human movement behaviour. \( \tau \) assumes possibly different state transitions for different environmental conditions; therefore it can be imagined as using a different VAR-iHMM structure to model data corresponding to each possible value of \( \tau \), while still sharing the same state parameters across all structures. For example, if \( \tau_t \in \{1, 2\} \) representing off-peak, and peak times respectively, the model can better identify the differences in behaviour between these two time regions while still allowing for similarities and overlaps in patterns should they occur. The final formalization of the model can be summarised with the following equations

\[
H \sim DP(\mu, Q), \quad G^{(i)}_j \sim DP(\alpha, \gamma), \quad G^{(i)}_k \sim DP(\tau, H), \quad \phi \sim Q. \tag{1}
\]

Where \( Q \) is the base probability measure, \( H \) is the master probability measure associated with the values of \( \tau \), \( G^{(i)}_j \) is the global probability measure for \( \tau = i \), and \( G^{(i)}_k \) is the random probability measure for state \( j \) when \( \tau = i \), \( \phi \) are the parameters of state \( k \) and \( \mu, \gamma \) and \( \alpha \) are the concentration parameters for each respective Dirichlet Process (DP). Figure 1 shows the results obtained from applying the model on taxi GPS data. The dataset records the journeys of multiple taxis over a period of 1 week in the city of Beijing, China as they conduct their day to day driving patterns. Beijing is highly congested and busy throughout most hours of the day, therefore, inner-city driving patterns will follow similar behaviour trends from morning until the late evening times. The clustered results confirm the hypothesis and identify various other behaviour trends as explained in the figure caption.

![Figure 1](image)

Figure 1: State 7 (Orange) represents observations that mostly take place in the center of the city starting around 9am when congestion peaks. The median velocity is 5.93 m/s. States 15 (green) and 23 (purple) which are also focussed in the center of the city with median velocities 6.74 and 8.38 respectively, gradually take the place of state 7 near midnight mimicking expected traffic trends of vehicular behaviour in lively urban cities. State 23 also overlaps with motorways along with state 22 (red) where the travel pattern differs when compared with inner city driving during congestion.